THE DISSERTATION COMMITTEE FOR RAHEL CHRISTINE KAHLERT CERTIFIES THAT THIS IS THE APPROVED VERSION OF THE FOLLOWING DISSERTATION:

RANDOMIZED CONTROLLED TRIALS TO EVALUATE IMPACT: THEIR CHALLENGES AND POLICY IMPLICATIONS FOR MEDICINE, EDUCATION, AND INTERNATIONAL DEVELOPMENT

Committee:

Peter Ward, Supervisor

Uri Treisman, Co-Supervisor

James Galbraith

Cynthia Osborne

Bryan Roberts

Randomized controlled trials to evaluate impact: Their challenges and policy implications for medicine, education, and international development

By

Rahel Christine Kahlert, MTheol; MPAff

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

DECEMBER 2012

Dedication

My work is dedicated to Faith and Steven who are happy that “mommy’s book is finished”, and to Robert who can’t wait to start his own.

Acknowledgements

I am grateful to all members of my dissertation committee—co-chairs Peter Ward and Uri Treisman, James Galbraith, Cynthia Osborne, and Bryan Roberts. As an interdisciplinary team, they have all provided me with their specific expertise, friendship and guidance. The eventual form of the dissertation most directly reflects my collaboration with my co-chair Uri Treisman whose optimism at times exceeded my own.

Ray Rist and Linda Imas-Morras generously allowed me to participate in the International Program for Development Evaluation Training. Howard White was instrumental in granting me access to the Network of Networks on Impact Evaluation meeting.

The International and Cross-Cultural Evaluation TIG of the American Evaluation Association created an environment for academic discourse that supported and informed my research.

Ryland Potter and Eric Abdullateef were reliable research companions.

Lynda De Jong and Talitha May were prompt editors. Jessica White-Sustaita was persevering in editing out my “Germanisms.”

RANDOMIZED CONTROLLED TRIALS TO EVALUATE IMPACT. THEIR CHALLENGES AND IMPLICATIONS FOR EDUCATION POLICY AND INTERNATIONAL DEVELOPMENT

Publication No._____

Rahel Christine Kahlert, PhD

The University of Texas at Austin, 2012

Supervisors: Peter Ward, Uri Treisman

Abstract: Policy makers in education and international development have lately gravitated toward the randomized controlled trial (RCT)—an evaluation design that randomly assigns a sample of people or households into an intervention group and a control group in order to measure the differential effect of the intervention—as a means to determine program impact. As part of federal regulations, the U.S. Department of Education and the U.S. Agency for International development explicitly declared a preference for the RCT.

When advocating for adopting the RCT model as the preferred evaluation tool, policy makers point to the success story of medical trials and how they revolutionized medicine from Medieval charlatanry to a modern life-saving discipline. By presenting a more nuanced account of the role of the RCT in medical history, however, this study finds that

landmark RCTs were accompanied with challenges, Evidence-Based Medicine had rightful critics, and opportunistic biases in drug trials apply equally to education policy and international development.

This study also examines the recent privileged role of the RCT in education and international development, concluding that its initial promise was not entirely born out when put into practice, as the national Reading First Initiative exemplifies. From a comparative perspective, the RCT movements also encountered major RCT critics, whose voices were not initially heard. These voices, however, seem to have contributed to a swing of the pendulum away from RCT primacy back towards greater methodological pluralism.

A major conclusion of this study is that policy makers should exercise great caution when using RCTs as a policy evaluation tool. This conclusion is arrived at via considering RCT biases, challenges, and limited generalizability; understanding its interpretive-qualitative components; and broadening the overall methodological repertoire to better enable evaluations of macro-policy interventions.

TABLE OF CONTENTS

Table of Tables

Table of Figures

CHAPTER 1: INTRODUCTION

1. The randomized controlled trial as privileged policy tool

Policy makers in education and international development have lately gravitated toward the randomized controlled trial (RCT) in order to determine policy impact. In 2005, the U.S. Department of Education (USDOE) officially proclaimed that RCTs are best for determining program effectiveness when evaluating federally funded education programs (U.S. Department of Education, Federal Register, January 25, 2005). Similarly, the U.S. Agency for International Development (USAID) stated in their new evaluation policy that RCTs generate the strongest evidence for impact evaluations, i.e., for determining the effectiveness of development programs (U.S. Agency for International Development, January 2011, 4). These two instances illustrate that the push for RCTs as a privileged evaluation tool became part of federal regulations and thus part of the policymaking process.

The RCT is a simple design that randomly divides a sample of humans (or other discrete units) into two groups: a treatment group that receives a new intervention, and a control group that does not receive the intervention. If the observations are large enough, randomization distributes the characteristics (i.e., extraneous factors) evenly across the two groups, thereby averaging potentially influencing factors. The RCT measures the difference in the average outcomes of these two groups, and this difference can be attributed to the intervention—i.e., it indicates whether the intervention is effective.

When calling for the RCT as the preferred evaluation tool in education policy and international development, RCT advocates point to the RCT's success story in the field of medicine, citing the following points: since the 1940s, medical researchers were able to find cures for deadly diseases such as Tuberculosis and Poliomyelitis with the help of RCTs; in 1970, the federal government was right in making the RCT the required standard before any new drug was approved to be put on the market; and in the 1990s, Evidence-Based Medicine brought the RCT to the medical practitioners' attention. This

focus on the RCT had helped to revolutionize medicine from a Medieval charlatanry to a modern life-saving science.

The policy makers' hope has been that the RCT would equally revolutionize education policy and international development, as it did medicine: that it would bring light into the uncertainty surrounding which interventions in educational and development are effective; that the RCT findings would adjudicate policy disputes on age-old policy questions; and that privileging the RCT would bring education and international development out of the Dark Ages and transform the two fields into scientifically based, modernized policy fields.

The policy problem: RCTs surrounded by misconceptions

When advocating for adopting the RCT model as the preferred evaluation tool, policy makers point to the life-saving history of medical trials and thereby make two key assumptions: First, that the RCT is indeed the ideal, unchallenged evaluation approach in medicine; second, that the RCT model could be directly transferred from medicine to education policy and international development. However, these assumptions are not without problems, as I show in following chapters. The policy problem is that misconceptions surround the RCT as policy tool in education and international development. Given these misconceptions, the RCT becomes a tool of belief and legitimization rather than science.

Why the policy problem matters

The privileged use of an evaluation approach has potentially far reaching policy consequences. Based on evaluation results, policy makers may start a new intervention or stop an existing one; or they may decide how much funding to allocate for a particular program in the next funding cycle. The fact is that privileging an evaluation approach, such as the RCT, necessarily devalues other evaluations approaches.

The shift toward RCTs as the privileged evaluation tool may lead to unintended consequences in policy making. The focus on RCTs would unjustifiably skew the type of possible policy interventions being evaluated, bypassing other, potentially more

promising, policy interventions. For the RCT can only be applied to certain types of policy interventions that allow for randomization—i.e., ones that involve a large number of observations, amenable to the construction of a control group. Certain policy solutions, including macroeconomic or environmental policies, cannot be randomized and thus could not be evaluated. The knowledge gained by RCTs alone may result in biased policymaking—one of the very problems the adoption of the RCT model was intended to address in the first place.

Research problem

RCT advocates in education and international development praise the success of the RCT model in medicine and call for adopting it in order to make their policy field more scientific. The research problem is that policy makers do not possess a nuanced understanding of the RCT as an evaluation tool for policymaking: first, they need to understand what the role of the RCT model was in medicine and what challenges it faced; and they need to understand what additional challenges arise with importing the RCT model into other policy fields like education and international development. Ignoring these challenges may result in uninformed decisions. Non-reflective overreliance on RCTs may lead to biased policy solutions.

This research constructs a more nuanced understanding from a cross-disciplinary examination of RCT use and RCT debates across policy fields. For example, although there has been research on the challenges of RCTs in medicine, no research has considered how these challenges and lessons of medical RCTs translate to the fields of education and international development.

Research questions

In order to determine the proper role of RCTs across three fields of policy-making—medicine, education, and international development—I pose the research question: What lessons for program evaluation can be discerned from the use of the RCT model in these three distinct areas of public policy making, and how can these lessons inform public policy evaluation? To answer this larger question, I first address the following sub-questions:

1. How did the RCT model emerge and develop in the field of medicine?
2. How did the RCT model get imported into the fields of education and international development?
3. How did the three policy fields deal with critics of the RCT model?

I first examine how the RCT approach emerged in the field of medicine. Medicine is of special interest because it embraced and institutionalized the RCT approach first and served as a model for modernizing the other two policy disciplines, education and international development. In the fields of education and international development, I analyze how the RCT model gained renewed interest, and how RCT advocates encountered and dealt with opposing views on the role of the RCT model.

Based on my prior analysis, I compare the RCT model's rise to prominence and its criticism across the three policy fields, and I show that the "RCT pendulum" is swinging away from RCT primacy back towards methodological pluralism. I compare the arguments of RCT advocates and RCT critics in order to provide a more nuanced understanding of the RCT as an evaluation tool. Lastly, I identify challenges of using the RCT model across the three policy fields, taking into account their respective differences. A major purpose of this study is to propose lessons and create recommendations for future work in impact evaluation in public policy—which ties back to the larger research question.

Contribution

My study provides a framework for a more productive discussion surrounding methodological choices for impact studies. It initiates an interdisciplinary dialogue on the policy role of impact studies, and RCTs in particular—specifically, its uses and challenges. I show, for instance, how different stakeholders refer to very different concepts when using terms like "impact" or "RCTs" in discourse. By comparatively analyzing the RCT use and debates across three policy fields, I contribute to the transfer of lessons from medicine to education and international development when it comes to the RCT approach. For example, the problem of "inclusion and exclusion criteria" in

drug trials translates to the question of sample heterogeneity in education and international development. I find that policy makers, researchers, and evaluators typically argue from their professional frameworks to make their case. For example, economists in international development use their quantitative framework of science to advocate for the RCT model, but they do not ask whether it takes into account the realities of program implementation.

The example of the Network of Network on Impact Evaluation (NONIE; cf., chapter four) contributes to understanding semantic and epistemological perspectives that guide methodological choices. I surveyed the extant literature about and produced by NONIE, and I collected primary data on the NONIE process in the form of observation notes. I mastered the literature on the larger context of the debates on impact evaluation in international development. Excepting Alexandra Caspari's policy work for the German Ministry for Economic Cooperation and Development (2008) and Howard White's internal view of the NONIE debates (2010), I am not aware of any use of NONIE working papers outside of my analysis.

Organization of study

My study is organized into five chapters. In the current chapter (chapter 1), I explain my methodological framework, a hermeneutic approach, and apply it to the rhetoric of two RCT advocates on importing the RCT model from medicine to education policy and international development. In a brief analysis of a school-based deworming RCT, I then demonstrate how RCT rhetoric could influence a shift in funding focus, despite unresolved policy issues. Lastly, although RCT advocates mostly use medicine as a reference point, I show that the origin of the RCT theory lies in agricultural statistics and in Ronald A. Fisher's work in particular. My analysis of Fisher's theory helps in understanding the challenges of RCTs, from which other policy fields such as medicine, education, and international development may equally benefit.

Chapters 2, 3, and 4 encompass the individual analyses of the RCT model in each of the three distinct policy fields. These chapters begin by analyzing several examples of RCTs, followed by examining the RCT movements in each field and arguments of its critics.

In chapter 2, I analyze the RCT model in medicine from several angles. First, I analyze two landmark RCTs, the Streptomycin trial on Tuberculosis in 1948 and the Salk trial on Poliomyelitis in 1954, because RCT advocates in education policy and international development point to these cases as successes. I show that even these trials faced challenges and provide lessons for policy. I then examine how the RCT became the regulatory standard in the drug approval process in 1970--though not without hurdles—and how Evidence-Based Medicine brought RCT primacy to the clinician's office in the 1990s. Finally, I analyze countermovements in mainstream medicine. In particular, personalized medicine and comparative effectiveness research argued for a more inclusive evidence base for effective interventions, moving beyond (quasi-)experimental knowledge generation. I show that even in the field of medicine, the RCT model faced several challenges and was always accompanied by criticism—criticism that could and should inform other fields.

In chapter 3, I analyze the RCT model in U.S. education from various perspectives. First, the Tennessee Class Size Reduction experiment in the 1980s became the poster child for successful RCTs in education. The California Class Size policy, however, illustrates the challenges in transferring RCT findings to different educational and policy contexts. Second, I analyze the National Reading Panel's work, which attempted to utilize and adapt the medical model for education policy. The subsequent scientifically based reading policy experienced implementation problems at the state level and steered school districts towards corporate, rules-based practices rather than towards evidence-based practices. Next, I show how the No Child Left Behind's Reading First initiative affected micropolitical decision making which resulted, for example, in regulatory alignment rather than a best practice in textbook adoption. Last, an illustrative case details the Federal Priority of RCT evaluation methods in education, which triggered many negative responses from the community of education evaluators. Although these were not acknowledged at first, the pendulum of RCT primacy seems to be swinging back to greater methodological pluralism.

In chapter 4, I analyze the RCT model in international development from several angles. First, I review the illustrative case of Progresa, a Mexican conditional cash transfer program, which had an RCT attached from its founding. The rigorous evaluation effort may have helped the program survive political leadership changes in Mexico, although the actual program impact seemed smaller than anticipated. Despite widespread RCT evaluations, conditional cash transfer programs still pose many open questions regarding their design and implementation. Second, I analyze the Center for Global Development's report on rigorous impact evaluations, which roused the international development community and led to an intense debate about how to best evaluate development impact. Third, I examine one such debate that took place within the Network of Networks on Impact Evaluation (NONIE), a multilateral donor network. The members of NONIE struggled to produce a guidance on impact evaluation, which should have defined what a high-quality impact evaluation would entail. Voices within NONIE criticized any hierarchical thinking (e.g., the RCT being the top choice) when making methodological decisions. Yet a compromise could not be reached.

In chapter 5, I provide a comparative analysis along two dimensions; first among the RCT movements in the three policy fields of medicine, education, and international development; and second, among the arguments between RCT advocates and RCT critics. I find that terminological differences of key terms, including the term "randomized controlled trial," have convoluted the arguments within the RCT debates. Epistemological differences prevail regarding what science is. Nevertheless, all sides agree that a triangulation of different methodological approaches and tools would increase the quality and relevance of an evaluation.

Finally, I show how the RCT debates in the fields of medicine, education, and international development can and should inform evaluation policy going forward. By comparing RCT challenges across fields, I argue that policy makers need an awareness of the challenges of the RCT model in order not to over- or understate RCT findings. Policy recommendations include understanding RCT biases, qualitative elements, and representativeness. Furthermore, executive summaries of policy experiments should caution about the limitations of the RCT design and its implementation. Scientific

reviews of evaluation findings can help in their utilization in the political decision-making process. Based on these recommendations, impact evaluations can obtain the necessary level of influence on policy decisions. The hope is that constructive debates further the quest for methodological rigor, quality, and relevance in the evaluation field. This can lead to both a stronger evaluation profession as well as better public policy choices based on methodologically sound evaluations.

2. Methodological and conceptual framework

My study uses the methodological approach of text interpretation in the hermeneutic tradition. Simultaneously, hermeneutics serves as a conceptual framework for my study because it situates the perception of RCTs with respect to the natural sciences. When RCT advocates praise the RCT model as the best evaluation tool, they typically harbor a perception of the natural sciences as being an objective and infallible means for uncovering truth. Hermeneutics enables me to demonstrate that privileging the RCT model is an attempt to model policymaking after the natural ("hard") sciences, and more importantly, that this approach comes at a price.

Natural sciences and the humanities, and where the RCT approach goes

Hermeneutics has generally been concerned with the interpretation and understanding of texts. In the 19th century, hermeneutic text interpretation assumed a strong historical-critical focus, combined with an emerging self-understanding of the humanities as distinct from the natural sciences. In his "Studies toward the foundation of the human sciences" (1970), originally published as "Abhandlungen zur Grundlegung der Geiseswissenschaft" in 1924, Wilhelm Dilthey laid the foundation for clearly distinguishing between the natural sciences vs. the humanities, which included history, political economy, law, philosophy, and the arts.[1]

Dilthey conceived of the natural sciences as the "explaining sciences" and the humanities as the "understanding sciences" or "hermeneutic sciences." The natural scientist was

[1] The term "humanities" and "human sciences" are only approximate translations of the term "Geisteswissenschaften", which more literally means "sciences of the mind."

concerned with describing and explaining the connections and causations of natural phenomena via experimentation. Experimentation, in Dilthey's worldview, referred to interventions performed and measured by a natural scientist in order to establish and test hypotheses about the natural world.[2] Conversely, for Dilthey, the human scientist was concerned with understanding "human beings, their relation to one another and outer nature" (Dilthey, 2002, 91). These human phenomena are fundamentally more complex and less uniform than the natural phenomena. These would be phenomena of an "immeasurable" reality (Dilthey, 2002, 142), or at least a reality not as readily measurable as physical phenomena. Subjects of study could vary in scope from individual expressions, social organizations, historical movements, or nations. The RCT movements are likewise subject to the principles of the human sciences. All of these subjects hold in common their "human-socio-historical reality" (Dilthey, 2002, 103). While the natural sciences were concerned with the discovery of universal laws behind the natural phenomena, the human sciences focused on understanding concrete expressions of the human productions (Dilthey, 2002, 103). Sometimes, the natural and human sciences overlapped. The study of language, for instance, includes the physiology of speech organs just as much as the theory of meaning and the sense of sentences (Dilthey, 2002, 103).

Dilthey observed that the methodological approaches of the humanities had traditionally been based on the natural sciences. One reason for this dependence was that the natural sciences had formalized their methods—including the comparative and experimental methods—first (Dilthey, 2002, 152). Dilthey criticized the human scientists for misguidedly attempting to legitimize their work by adapting natural science approaches:

> Even today [1924], when psychologists, educationalists, linguists, and aestheticians tackle specific problems, they will often ask themselves whether the means and methods for the solution of analogous problems in the natural sciences can be fruitfully applied in their own field. But despite such particular points of contact, the procedures of the human sciences are from the beginning to the end different from those of the natural sciences" (Dilthey, 2002, 152).

I make a similar argument that the adoption of the RCT approach in the human and social sciences is an attempt to establish them as rigorous sciences, so that they will enjoy the

[2] Note that Dilthey did not equate experimentation with RCT yet.

same perception and privileges that the established natural sciences enjoy. The RCT is a substitute for the laboratory of the natural sciences. An RCT approach determines an intervention's effect, but it falls short in understanding its underlying processes. The medical field widely used the RCT approach in order to determine drug effects on human biology. However, when used as an evaluation tool in policy fields of education or international development, the RCT model is transferred to disciplines of the humanities, which brings both gains and costs. Gains include greater clarity about whether a certain historical intervention achieved certain effects; costs include the extreme focus on specific information, which does not always allow for understanding the causal mechanisms of an intervention—an issue central to the humanities and thus hermeneutics.

I argue that the adoption of the RCT model in public policy is an attempt to make policymaking credible and its programs legitimate, because it gives the impression of transforming policymaking into a hard natural science. One of the major questions underlying this study is where public policy, with its unique methodological approaches, fits into the canon of sciences. I argue, as Dilthey did for the humanities, that modeling public policy after the natural sciences may increase their scientific recognition, but it comes at the expense of reduced understanding and thus reduced policy relevance.

How the hermeneutic process works

The German philosopher Hans-Georg Gadamer universalized the concept of hermeneutics to human communication and reasoning in general. In "Truth and method" (2004), originally published as "Wahrheit und Methode" in 1960, Gadamer laid out a system of interpretation and understanding as a fundamental means by which humans communicate with each other.

Gadamer explained the process of understanding as follows:

> The process of construal is itself already governed by an expectation of meaning that follows from the context of what has gone before. It is of course necessary for this expectation to be adjusted when the text calls for it. This means, then, that the expectation changes and that the text unifies its meaning around another expectation. Thus the movement of understanding is constantly from the whole to the part and back to the whole. Our task is to expand the unity of the understood

> meaning centrifugally [in konzentrischen Kreisen]. The harmony of all the details with the whole is the criterion of correct understanding. The failure to achieve this harmony means that understanding has failed (Gadamer, 1960, 275; Gadamer, 2004, 291).

The interpreter starts with a proto-understanding, which is shaped by her historical, socio-political context. It may consist of misunderstandings and preconceived prejudices, based on the historical and cultural distance between the interpreter and the texts (Gadamer, 1960, 173). This proto-understanding guides the initial interpretation, which she then revisits and reinterprets in the light of the whole text and new text passages. The interpreter attempts to continuously explicate the tension and bridge the distance between herself and the texts. The interpreter oscillates between her own understanding, the text's details, and the text's whole. Gadamer's "criterion of correct understanding" is whether the details are in harmony with the whole. Gadamer referred to this circular process of reinterpretation as the hermeneutic circle.

I counter that Gadamer's idea of the "harmony with the whole" is an ideal type, because a text and its details may not always appear univocal. Documents produced by committees, for instance, may reflect a diversity of voices. They may not be able to reconcile these voices. The perceived incoherence and ambiguity of the text could stem from both the inadequate understanding of the interpreter or the incoherent nature of the text. The interpreter's task is thus to distinguish between these two sources of incoherence. Therefore, the interpreter needs to look at the historical context of the text to determine possible factors of incoherence.

As I show in the example of the Network of Networks on Impact Evaluation (cf., chapter 4), different stakeholders collaborated on the production of draft documents for impact evaluation guidelines. The NONIE stakeholders' professional and cultural differences most likely affected how they contributed to and understood each others' texts and conversations. The task is to explicate these differences and to make sense of the, at times multivocal, texts.

My own hermeneutic process and motivation

In line with the hermeneutic circle, I have changed my perception of the debates on the RCT primacy over time. Before I started my research, I was an RCT skeptic. I was worried about the potentially negative side effects of the 2001 *No Child Left Behind Law's* insistence on "scientifically based research." I assumed that they operated under the prejudice of scientism. The Federal Priority on scientifically based evaluation methods in 2003 furthered my suspicion of the possibly detrimental effects of the scientifically based research focus, which privileged RCTs as evaluation approach in education. My suspicion surrounding RCTs was my original motivation for developing my research agenda around the RCT advocates and their critics.

Over the years, I have gained a much more nuanced understanding of the RCT model and have become more appreciative of the RCT as a policy tool, but also more pointed in my criticisms of its misuses. I learned to distinguish between the rhetoric of RCT "scientism" and the reality of RCT uses, where they are useful and where they are inappropriate. Although my original suspicion surrounding RCTs has become more founded by a better understanding of their limitations, I also conclude that RCTs make valuable contributions to policymaking, especially when accompanied by other approaches. Moreover, building bridges between the RCT model and social science approaches (including hermeneutics) can strengthen the value of RCT findings, leading to higher-quality evaluations and smarter policymaking. However, the limited applicability of the RCT tool in policy evaluation remains a valid concern. Their exclusive use would indeed bias policymaking.

Document collection

My study is based on texts of various types, depending on the particular policy field. In the discussion of RCTs in agriculture, I mainly relied on Ronald A. Fisher's texts from the 1920s and 1930s (cf., chapter 1).

Chapter 2 (on medical trials) relies on original source texts such as Austin B. Hill's writings on randomization in the 1930s; published trial results (Streptymycin Trial on Tuberculosis; Salk Poliomyelitis trials); legislative texts and council reports on drug efficacy; Evidence-Based Medicine's original publications and those of its critics; and

speeches from federal commissioners in the United States and the United Kingdom on the new directions of federal policy.

In chapter 3 (i.e., RCTs in education), I analyze originally published trials (Tennessee STAR trial; California Class Size Reduction); meeting notes and publications from the National Reading Panel and its critics; legislative texts and hearings; federal registers; and commissioner reports. I also use electronic lists of the American Evaluation Association (Evaltalk) to analyze the evaluation-internal debate on the Federal Priority.

In chapter 4 (i.e., RCTs in international development), personal communications were instrumental in collecting what has been called "grey literature." This term refers to either unpublished or electronic documents, rather than formally published documents. I collected materials from professional development trainings and professional evaluation conferences such as from the American Evaluation Association. For example , the International Program for Development Evaluation Training (IPDET), organized by the World Bank in summer 2007, allowed me to make the connections for attending the Network of Networks on Impact Evaluation meeting in January 2008—a meeting that became an illustrative case of the RCT debate.

I recorded all documents in Citavi, a bibliographic data management tool, which allowed for detailed summaries, quotations, and annotations; it also allowed for concurrent comparison of various, sometimes diverging sources. For example, I trace the RCT debate within the Network of Networks on Impact Evaluation (NONIE) via several documents: background papers, draft guidelines, visual presentations from various working groups, meeting notes, and personal communications. The citations in the bibliography are organized by their text format (eg., postings on electronic mailing lists, presentations, interviews, reports etc.).

3. The rhetorics of RCTs: From the Dark Ages to modernity

RCT advocates in education and international development use the modernization of medicine as an exemplary case of how a discipline shed its unscientific roots and became

a true science. They observe that adopting the RCT as the methodological standard for evaluating new drugs was largely responsible for the success of medical science and its life-saving practice. RCT advocates accuse the fields of education and international development of remaining in an unscientific stage—comparable to Medieval blood-letting or leech treatments. Doctors were convinced that their treatments worked; but had their convictions been scientifically evaluated, they would have found them to be at fault. RCT advocates in education and international development argue that in order for their fields to become equally successful and scientific as medicine, the adoption of the RCT as a matter of course will be necessary.

In the following, I choose one example from the field of education and one from international development to illustrate this quest for adopting the medical RCT model. Using a hermeneutic-interpretive approach, I find inconsistencies in the line of arguments—which sets the stage for a closer analysis of how the medical RCT model could inform education policy and international development.

The illustrative example in the field of education is Valerie Reyna's presentation at a conference organized by the U.S. Department of Education (USDOE). Reyna, a senior research adviser at USDOE, suggested utilizing the medical RCT model for education reform:

> We know on the basis of experience that anecdotes have turned out to be false and misleading. Sometimes they are very representative, sometimes they're not. The problem is we don't know when. There's an analogy to medicine that I have obviously drawn on already. The first example, of course, is the classic one of when they used to bleed people. People would get sick. You know, I think it was when George Washington was bled that contributed to his death […].
>
> The bottom line here is these same rules about what works and how to make inferences about what works, they are exactly the same for educational practice as they would be for medical practice. Same rules, exactly the same logic, whether you are talking about a treatment for cancer or whether you're talking about an intervention to help children learn. […] The reason I have the word "brain surgery" up there is that I think, you know, when we talk about medicine and things like brain surgery and cancer, it is very, very important to get it right. We all recognize that and most of us buy into that. You know, that you've got to have randomized clinical trials because we want to be able to benefit for these treatments for cancer. But when we teach students we really are engaging in a kind of brain surgery." (Reyna, presentation, February 2, 2002)

Reyna used the metaphor of education as "brain surgery" to illustrate the life-enhancing and possibly life-threatening nature of education, parallel to the nature of medicine. Reyna compared current educational practices to the outdated medical practice of bleeding, based on anecdotal, and thus unsubstantiated, evidence. Similar to bleeding patients, current education practices would harm students rather than help them. According to Reyna, the field of medicine had successfully chosen to be life-enhancing by adopting RCT standards. RCTs would distinguish between what works and what does not work. Reyna suggested a similar reform for the field of education by adopting RCT standards, modeled after the medical success story.

Note some inconsistencies in Reyna's arguments: First, the information about whether Washington had died from blood-letting was not based on RCT knowledge, but on historical reports. In fact, I do not know of any RCT that established that blood-letting is harmful to a human. Reyna referred to "experience," on which basis anecdotal information (e.g., that leeches heal the sick) turned out to be false. This experience itself, which Reyna used to make her argument, however, is not based on an RCT, but rooted in historical observations. There is a tension between privileging the RCT model on the one hand, and relying on observational information (e.g., in the case of Washington's death) on the other hand without acknowledging it as a reliable source of knowledge.

Second, using the metaphor of "brain surgery" to argue for RCT standards is not conducive to Reyna's line of argumentation either. Compared to drug testing, RCTs are difficult to use in brain surgery. Double-blinding, a core component of medical RCTs, is hardly justifiable due to ethical concerns. The fact that it is more difficult to use RCTs for brain surgery is an indication that educational "brain surgery" may be even more complex, because the procedures are not just biological, but also social and emotional. In sum, although Reyna's argument was that education policy should follow the lead of medicine in adopting the RCT as an evaluation standard, there are certain incongruencies in her argumentation.

Similar arguments about mimicking the medical RCT model can be found in international development. An illustrative example is Esther Duflo's talk in the Technology, Education, and Development (TED) series, with approximately 300,000 viewers. As co-director of the Jameel Poverty Action Lab (J-PAL), Duflo has advocated for the RCT in evaluating anti-poverty interventions in international development:

> "The thing is, if we don't know whether we are doing any good, we are not any better than the Medieval doctors and their leeches. Sometimes the patient gets better, sometimes the patient dies. Is it the leeches? Is it something else? We don't know. […] It is not the Middle Ages any more, it's the 21st century. And in the 20th century, randomized controlled trials have revolutionized medicine by allowing us to distinguish between drugs that work and drugs that don't work. And you can do the same randomized, controlled trial for social policy. You can put social innovation to the same rigorous, scientific tests that we use for drugs. And in this way, you can take the guesswork out of policy-making by knowing what works, what doesn't work and why. […] These economics I'm proposing, it's like 20th century medicine. It's a slow, deliberative process of discovery. There is no miracle cure, but modern medicine is saving millions of lives every year, and we can do the same thing." (Duflo, presentation, February 1, 2010)

Duflo compared today's international development interventions to Medieval leech cures. Doctors were wrongly convinced that the cures worked, based on individual observations. According to Duflo, most of the anti-poverty interventions in the 21st century were equally based on belief rather than knowledge. Thus, people assumed that they work, but they would never know whether they in fact did work. Next, Duflo argued that this situation could change via the RCT revolution. The RCT would finally enable distinguishing drugs that work that drugs that do not. Just as the RCT standard transformed medicine into a life-saving science, it has a potential for turning social policy into a scientific enterprise and to actually help people.

Although Duflo constructs parallels between education and social policy, she indicates a desire for RCTs to accomplish more in social policy: First, Duflo referred to medical RCTs that distinguish between drugs that work and drugs that do not work. Later, she wants social RCTs to "know[ing] what works, what doesn't work and why." Duflo added the policy-relevant question "why" a social intervention worked. As I conclude in the last chapter, RCTs alone cannot answer the "why" question, but need to be supplemented by other methodological approaches to do so. For drug evaluations, the "why" question is less relevant because drug interventions are rather discrete, whereas social interventions

are often multidimensional within a complex environment. The answer to why a social intervention works is important in order to transfer the findings to other contexts. Thus, Duflo's vision of saving millions of lives every year via RCT findings alone is less realistic for social than for medical interventions.

By referring to medical history, Reyna and Duflo proposed that the RCT become the privileged model for evaluating interventions in their respective fields. They hoped that the RCT model would help their fields modernize and arrive at scientific knowledge for answering long-awaited policy questions. Their hopes were probably more optimistic than realistic. The facts that Reyna used non-RCT knowledge to make her argument and that Duflo overlooked the inability of RCTs to answer policy-relevant "why" questions illustrate that rhetoric is probably further from the reality.

4. The rhetorics of RCTs put into practice: School-based deworming

In the TED talk cited above, Esther Duflo referred to the RCT on school-based deworming in Kenya, among others, as a successful example of informing social policy. According to Duflo, the RCT findings had established the following:

> And for every hundred dollars, you get almost 30 extra years of education. So this is not your intuition, this is not what people would have gone for, and yet, these are the programs that work. We need that kind of information, we need more of it, and then we need to guide policy (Duflo, presentation, February 1, 2010).

In the following, I take a closer look at the deworming RCT regarded as a success story in guiding effective social policy. This RCT is an example of how the creation of RCT evidence was combined with its promotion in the right circles, which led to a school-based deworming movement. Based on the RCT evidence, deworming interventions were scaled up in several African countries, and many donor countries have since committed additional resources to the eradication of intestinal worms. However, many questions remain unanswered, and it is therefore unclear whether "grasping at straws" with respect to this particular RCT justifies the exclusion of streams of non-RCT evidence. I argue that the limited RCT findings appealed to a particular audience, economists especially, but the finding may not justify the RCT's far reach.

The economists Edward Miguel and Michael Kremer conducted an RCT in a small part of Western Kenya from 1997 to 2001 (Miguel & Kremer, September 2001; Miguel & Kremer, 2003). 75 schools serving approximately 30,000 students were randomly assigned to treatment condition and control condition. The control schools received worm treatment two years later. Miguel and Kremer found that administering deworming medication to students at a school decreased student absenteeism by 7 percent in two years, from ca. 23 percent to 15 percent in the first year, and from 35 percent to 29 percent in the second year (Miguel & Kremer, September 2001; Miguel & Kremer, 2003). They concluded that, indeed, school-based deworming increased school attendance. They estimated that the cost per additional year of school participation was $3 and concluded that this treatment was far cheaper than alternative ways of boosting primary school participation (Miguel & Kremer, September 2001, 2). The researchers did not find educational gains based on test scores, which would have been a higher-level impact than school participation (Miguel & Kremer, 2001, 1).

Challenges of the school-based deworming RCT

First, although school-based deworming is a laudable intervention, it may not have actually needed randomized evidence. Medical RCTs had already established the effectiveness of the distributed deworming medications like Albendazole and Praziquantel, and they had determined their positive effect on health. Healthier children are more likely to have higher school attendance. The question is whether an additional RCT would be necessary for establishing positive educational outcomes by themselves.

Second, the deworming intervention reduced absenteeism where intestinal worms were a wide-spread problem. The researchers chose the two divisions of the district Busia in Western Kenya, with densely settled farming and high infection rates (Miguel & Kremer, September 2001, 7). In fact 92% of students were infected (Miguel & Kremer, September 2001, 10). Such an intervention would not work in many other countries due to minimal infection rates; in that case, alternative ways may indeed be far cheaper than this intervention.

Third, the cost estimate of 3 U.S. dollars per added school year seems like "hard data." Note that Esther Duflo also cited 30 extra years per hundred dollars, which sounds impressive, but most likely fluctuates based on contextual factors. In the case of West Kenya, absenteeism was still 35 percent in the second year despite the deworming medication, possibly due to the El Niño flooding. Factors such as flooding and drought might equally influence school absenteeism, regardless of whether students receive anti-worming medication. The intervention would also not work well for girls in countries where female education is not valued. Policy makers need to take into account the particular situation—e.g., if an area is affected by drought, flooding, or cultural beliefs—before analyzing costs and benefits.

Fourth, Miguel and Kremer did not use placebo pills in control schools. Therefore, the administration of the pill, rather than its active ingredient, could have increased school participation. For example, school participation may have increased due to parents' perception that the school cared for their child's health via the administration of these pills. According to Scriven, the lack of placebo in social-program RCTs could lead to major distortions (Scriven, 2008). Although, the deworming pills had been proven to be effective in treating intestinal worms, the simple fact of teachers administering pills could increase school attendance. Administering placebo pills at the control schools would have been unethical because the health-effects of the actual pills were already known, and a placebo would have deceived the participants. Furthermore, the low costs of the pills would not have justified excluding the other schools. Without using a placebo, however, the black box of school-based deworming has not been fully opened: What worked, the pill ingredients or the perception of the pill working?

The deworming movement and channeling of funds

Despite of the limited evidence of a small pilot, Michael Kremer and Esther Duflo succeeded in making the RCT findings known at the 2007 World Economic Forum Annual Meeting in Davos, Switzerland. As members of the Young Global Leaders Education Task Force, they launched the *Deworm the World Initiative* (DTWI) in Davos.[3] The organization based its mission on the claim that mass deworming at schools

3 http://www.dewormtheworld.org/?q=node/68, accessed Oct 11, 2011.

improved school participation, increased ultimate earnings and workforce participation, and would be an efficient and effective way to treat large numbers of children. The organization asserted that all of these claims were based on rigorous evidence. As DTWI stated on their website: "This evidence was a breakthrough. School-based deworming was globally recognized as a 'best buy' for development." DTWI garnered additional interest from funders, such as The United Nations (WHO, UNICEF), the World Bank, and several private-sector donors like the Gates and Dell foundations. Academic institutions, including J-PAL, are partners of DTWI. DTWI provided technical assistance to help the Kenyan government treat 3.6 million children in 2009 (MIT Technology Review, January 2010), and they promoted deworming in other developing countries.

The RCT evidence allowed Kremer to convincingly draw attention from economists to the findings, thereby initiating an international deworming movement that was supported by the private and public sector funds. Economists by profession trusted the provided RCT evidence, and they became the original actors of the school-based deworming movement. DTWI gained momentum, and others from the public sector joined the movement.

From a funding perspective, resources were channeled to the deworming cause. There are many other health problems in developing countries, malaria being just one example. Malaria, however, does not have an effective vaccine yet, and many additional resources are needed to create an effective vaccine. One major question is whether the deworming case generated additional funding or just redirected funding streams from, for example, discovering an effective vaccine for Malaria. An unintended side effect of funding interventions based on small-scale RCT evidence is that they could be crowding out other interventions—interventions that are potentially even more far-reaching.

Intestinal worms are also a symptom of the problem of unclean water. Other interventions may have also been more effective in treating the problem at its core. Reinfection could easily happen if school-based deworming was stopped, because the problem was not treated at its core. From this perspective, the success story of school-based deworming is losing some of its "best buy" qualities. A more comprehensive policy review might be

needed to justify the increased funding towards this intervention, regardless of RCT evidence.

This example illustrates how one single average data point from an RCT made history in international development. Its evidence created an expansive movement that believed in deworming in schools. It has been unclear, however, whether other interventions against intestinal worms (e.g., by prevention) or other intervention against absenteeism (e.g., school meals) would have been more effective policy solutions. Simultaneous RCTs in Western Kenya would have been necessary to make the results comparable and answer these questions. But even then, the RCT results would only apply to that particular context at that certain time and in that certain place. As I stated above, the RCT findings would remain ignorant beyond the duration and the location of the experiment, which also prevents any insights into the adaptation of the intervention to related contexts. Therefore, to take a closer look at the RCT and to understand its potential role in policymaking, I am going back to its theoretical foundation. This foundation does not lie in medicine, but in agriculture.

5. Theoretical foundation of the RCT model: R.A. Fisher's fertilizer studies

RCT advocates in education and international development cite medical trials to make their case for using the RCT model in their field. Although medicine is currently the stronghold of RCTs, the theoretical foundation of the RCT dates back to the agricultural statistician Ronald A. Fisher in the 1920s. In his fertilizer studies, Fisher recognized that random assignment of soil plots would tackle the problem of soil heterogeneity due to innumerable causes, which had made it difficult to compare soil plots in the past.

In the following, I review Fisher's work on the RCT, citing original journal articles and books from the 1920s and 1930s (cf., bibliography) in order to understand the theoretical foundation of the RCT model. Going "ad fontes" is a key principle of hermeneutics in understanding the original intent.

Backdrop of the RCT theory: Fisher's quest for unifying the sciences

Conceptually, Fisher stood in the British tradition of biometrics spearheaded by the biological statisticians Francis Galton and Karl Pearson. Using statistical methods, their goal had been to raise biology to the "status of a more exact science" (Galton, 1901, 10). Their journal *Biometrika*, founded in 1901, united their efforts to foster statistical "study of differences" in biological phenomena (Biometrika, 1901, 1). For Fisher, biometrics was part of an "intellectual liberation," similar to the discovery of geometry in antiquity (Fisher, 1932).

Fisher hoped for statistics to become the unifying foundation of the different sciences. He subscribed to the "efforts to unify the theoretical concepts underlying the two great branches of human knowledge" (Fisher, 1932, 3)—the natural and the social sciences. Fisher took a "statistical view of the world" (Fisher, 1932, 11), where the idea of variation and the concept of chance were applied the same way to the expansion of gases, the human body, and human societies. The statistical view would study these phenomena not as specific gas particles or human individuals, but as populations with inherent variations. In this vein, Fisher argued that: "Statistical methods are essential to social studies, and it is principally by the aid of such methods that these studies may be raised to the rank of sciences" (Fisher, 1925, 1).

The RCT theory grounded in fertilizer studies

In 1919, Fisher was appointed as statistician at the Rothamsted Experimental Station, the major center for agricultural research in the United Kingdom (Yates & Mather, 1963, 92). Such experimental stations had been established in the nineteenth century under the British law of Land Grant Colleges to institutionalize agricultural science as an academic discipline (Armitage, 2003, 925). The field settings of the research farm allowed for the experiments in a natural environment instead of a controlled laboratory setting. Fisher's task was to scientifically investigate the impact of fertilizers on crop yield.

Fisher found that plots naturally differed in crop yields up to 30 percent due to soil heterogeneity. For Fisher: "the greatest source of error in field experimentation is that due to the heterogeneity of the soil" (Fisher, 1931, 11). Fisher attributed variations in wheat

yield to annual variations caused by weather (e.g., light, temperature, moisture), soil deterioration, and other slow changes such as weed growth (Fisher, 1921, 108). The researcher could not make an exhaustive list of possible "causes of disturbances," as the "uncontrolled causes which would influence the result are always strictly innumerable" (Fisher, 1935, 21).

In his paper "The arrangement of field experiments" (1926), Fisher used the example of using manure for increasing crop yield to illustrate the problem of soil heterogeneity. The agricultural researcher had used similar seeds and treatments for two acres of land, with the only difference being that manure was applied to one acre but not the other. Even if he were to find a 10 percent difference in crop yield, the question was still whether the manure caused this difference: "What reason is there to think, even if no manure had been applied, the acre which actually received it would not still have given the higher yield?" (Fisher, 1926, 504) According to Fisher, the two plots may have had different soil composition and thus different fertility in the first place. The researcher would need to prove that the plot would have provided similar yields—an impossible quest; 500 years of comparative data points would have been necessary to do so reliably. Fisher suggested a much more efficient procedure to ensure comparability of the plots: Dividing each acre into 32 or 40 individual plots, pairing adjacent plots, and then randomly assigning these plots to treatment with or without fertilizer. As a result, the soil fertility and other factors that may influence crop yield would be equally distributed between fertilized and unfertilized plots (ibid., 505–506). The heterogeneity of the soil plots could be statistically controlled without requiring the physical isolation of the laboratory. After the treatment with fertilizer, the plots yielded a certain number of bushels of wheat. The difference between the average yield of fertilized and unfertilized plots would be the net impact or added benefit of the fertilizer.

Note that the unit of randomization and analysis are plots, i.e., areas of land. The size of the plots could vary, and an acre could be subdivided in different ways, such as in squares (e.g., 5 by 5 or 6 by 6) or strips (e.g., 4 by 8 or 5 by 8). Smaller plots yielded more units of analysis per acre and therefore made the experiment stronger, with higher degrees of freedom. The plot size per acre also depended on the type of farm machinery used and

the precautions against edge effects.[4] Proximity was an important concept in randomizing plots. Fisher pointed out that more proximate plots were more likely to be similar. Therefore he suggested a two-step process: First, creating blocks of adjacent plots, and second, randomly assigning those plots to treatment and control group. The randomization of adjacent pairs of plots would be advantageous as it would reduce the standard error between the units and therefore would yield more precise estimates (Fisher, 1926, 507). The researcher would not need to know the soil characteristics of those adjacent plots to assume their similarity.

Fisher made two arguments for random assignment: one being the more theoretical argument for guaranteeing the validity of significance testing, and the other being the more practical to reduce experimenter's bias. First, randomization was the "physical basis of the validity of the test of significance" (Fisher, 1935, 20), where the results would be governed only by the "laws of chance" (ibid., 20). Fisher recommended the probability of five percent as the standard level of statistical significance as a "convenient convention" (Fisher, 1935, 16), where one in twenty trials would yield results by chance coincidence.

Second, Fisher found that by random assignment, the researcher would not be able to "cook" the arrangement to suit his preconceived ideas (Fisher, 1926, 509). The traditional practice of purposeful assignment had relied on such preconceived ideas of what plots were comparable and allowed experimenters to control what unit went into the treatment and control groups. This practice, however, would have compromised the results, leading to overestimation or underestimation of errors.

Discussion: Fisher's RCT theory

Fisher's original theory of randomized experiments gives rise to several considerations and insights for using the RCT model. Whereas Fisher used examples from the field of biology and agriculture, in which RCT design and application were most advanced at his time, he also argued that the principles would be applicable to other fields, such as the medical and social sciences (Fisher, 1935, 11). Because Fisher was mainly familiar with

[4] Edge effects could impact adjacent plots resulting in possible contamination (Birk, 2005, 91).

agricultural and animal trials, he had, for example, little to say about allocation concealment, perception bias, or dropout rates, when dealing with human subjects.

First, Fisher's conceptualization of the sciences is juxtaposed to the hermeneutic thinking by Dilthey and Gadamer, as outlined above.[5] The hermeneuticists insisted on distinct methodological approaches for the natural sciences and the humanities, based on their different nature and goals of investigation (i.e., to explain versus to understand). On the contrary, Fisher wanted statistics to unify the natural and social sciences and, in fact, to propel the social sciences into the rank of a true science. By applying the RCT model to the social sciences and the humanities, researchers attempted to generate true science. This came at a price. The focus on averages between groups would lead to an extreme "reduction of data," as Fisher recognized (Fisher, 1925, 1).

Second, Fisher's own reflections illustrate how many more considerations went into the design and interpretation of an RCT than just purely statistical expertise. Despite the fundamental disagreement on science, Fisher and the hermeneuticists shared an important insight: Even natural scientists or experimentalists need to be critical of their work. Fisher called for a statistician's general intelligible ability. Dilthey and Gadamer would point to natural scientists' need for using hermeneutic skills. Fisher cautioned against the non-critical use of experiments. Statistical skills alone were insufficient for designing and interpreting experiments. Fisher distinguished between a statistician's technical craft, where he had special authority, and the craft of scientific inference, which would require general intelligible ability (Fisher, 1935, 2). A randomized experiment was based on inference like any other type of knowledge generation. Fisher assumed that it was "possible to draw valid inferences from the results of experimentation" (ibid., 4), and that it was possible to argue from consequences to their causes and from observations to their hypotheses. However, caution was warranted.

Fisher found that some uncertainty existed in the inference process from an event to its possible causes, despite its rigor. Fisher argued that a statistician's task was to determine

[5] Note that Fisher and Dilthey lived concurrently, though they most likely did not know each others' work due to language differences and geographical distance.

"how to evaluate the limitations of the data in hand" and to recognize the defects of the experimental technique (Fisher, 1933, 46). Fisher emphasized that every experiment should start with an explicitly formulated hypothesis, which might or might not be impugned by the result of the experiment (ibid., 19). This hypothesis could never be proved, however, but it could possibly be disproved in the course of experimentation. Fisher pointed out moreover that the selection of a hypothesis was always inductive, and thus preliminary to any deductive discovery (ibid., 6). However, inductive inference was the only process of knowledge generation (ibid., 8). Fisher argued that experimental observations had inductive elements and were directly linked to the existing body of knowledge (ibid., 9), as opposed to the purely deductive reasoning of geometry.

Third, Fisher argued that randomization would ultimately guarantee the validity of statistical significance testing. Fisher picked the *p* value of .05 of a test of statistical significance as a "convenient convention" (Fisher, 1935, 16), not as a rigid standard. However, the *p* value became a fundamental indicator of whether an intervention is effective (Hubbard & Lindsay, 2008, 70). The psychologists Hubbard and Lindsay observed that researchers had been overly reliant on the *p* value as an objective, useful, and unambiguous measure of evidence in hypothesis testing (Hubbard & Lindsay, 2008, 71).

Instead, researchers should display more caution in using this measure in testing hypotheses. For example, sample size would influence the meaning of the *p* value. When the sample size is large enough, almost any null hypothesis would have a tiny, statistically significant *p* value (ibid., 75). Peter Freeman illustrated this concern with hypothetical medical trial results, where the subjects received treatments A and B and were asked about their preference. In Table 1, the *p* value is .041, i.e., statistically significant in all four trials (Freeman, 1993).

TABLE 1: Statistically significant trial results with a *p* value of .041

Trial	No. preferring A	No preferring B	% preferring A
1	15	5	75.0
2	114	86	57.0
3	1,046	954	52.3
4	1,001,455	998,555	50.07

The preference rate in trial four of 50.07 percent would be considered equally statistically significant, as compared to the preference rate of 75 percent in trial one, in which the main difference is number of participants. This hypothetical trial illustrates that the *p* value is not sufficient in providing evidence for the effectiveness of a treatment. The test of statistical significance does not address the size of the effect, which may be a more relevant measure for decision makers.

Fourth, gases, the human body, and human societies may all seem to behave in similar ways, but one fundamental difference is human self-reflection and spontaneity, which influence the experimentation process.

Fifth, for RCTs outside the agricultural field, however, the concept of geographical proximity and the notion of partonomy were less relevant. The units of treatments are not geographically fixed plots or stuff type, but whole objects, i.e., human individuals and human groups, who are geographically mobile. Thus the principle of geographical proximity does not hold. Also, human self-reflexivity adds a new feature to the RCT model where individuals are influenced by their participation in RCTs and might change their perceptions or drop out of the RCT altogether.

Sixth, Fisher emphasized that no isolated experiment, however significant in itself, could suffice for the demonstration of a natural phenomenon and its cause (Fisher, 1935, 16). At least in theory, the problem of soil heterogeneity could be overcome by replication, by diminishing experimental errors and by providing the magnitude of those errors (Fisher, 1931, 12). Recall that the school-based deworming movement was based on one isolated RCT in Eastern Kenya, without replications in different contexts. Fisher would not have approved this use of a single RCT.

As I discuss in chapters three and four, this problem of selection bias holds true for other policy areas such as education and international development. The following section applies Fisher's concept of random assignment to the area of medicine, where allocation concealment and selection bias were major issues (Chalmers, 2001, 1162).

CHAPTER 2: THE RCT MODEL IN MEDICINE: THE REFERENCE POINT

Many stakeholders in education and international development view the dominance of the RCT approach in medicine as a desirable goal for their own discipline. Many view RCTs as having transformed medicine from Medieval charlatanry to a modern science and having thus significantly improved the welfare of the world. They view landmark studies, such as the Tuberculosis trials in 1940s and the Salk Poliomyelitis trials in the 1950s as major turning points for medicine, and they see the role of the Food and Drug Administration in the 1960s and 1970s as pivotal to making the RCT standard a reality. RCT advocates applaud the Evidence-Based Medicine movement in the 1990s for bringing the RCT-led decision making into clinical practice.

In what follows, I present a more nuanced account of these medical landmark trials and developments. My purpose is to help policy makers better discern appropriate lessons for using RCTs in education and international development. In the first section, I investigate the following: the theoretical foundation of medical RCTs by Austin B. Hill in the 1930s; the Streptomycin trial on Tuberculosis, which was the first formally published and still frequently cited RCT by the U.K. Medical Research Council in 1948 (Hill was the lead medical statistician); and the 1953 Salk Poliomyelitis trials in the United States. Although these trials led to a wider acceptance of RCTs in the medical sciences, I illustrate that they also faced ethical, logistical, and interpretive difficulties, which are often overlooked when citing them as success stories.

In the second section, I analyze the legal institutionalization of the RCT in the United States by the Food and Drug Administration in order to guarantee safe and effective pharmaceuticals for public distribution. The RCT as the sole means of drug evaluation was ultimately decided not by Congress, but by the federal court system in 1970. Despite its legal institutionalization, medical practice lagged behind medical science. The evidence-based medicine movement of the 1990s eventually brought the RCT requirement into the practitioner's office. The hierarchy of methods and its channeled

decision-making model became an exemplar in other fields such as education, social policy, and international development.

At the same time, calls for a widening of methodological approaches arose within the medical field to overcome challenges associated with RCTs. These challenges included limited external validity, black-box evaluations, and impersonalized decision-making. In the third section of this chapter, I analyze these responses within the medical field, including comparative effectiveness research and personalized medicine. Both responses argued for the widening of the evidence base and for tailoring medical research to individual patient's needs.

I show that the so-called success story of the RCT in medicine in fact exhibited many challenges, which RCT advocates in education and international development rarely cite. These challenges, however, are worth exploring because they are often more severe in the fields of education and international development, given that these require often more complex interventions than administering a drug.

1. The Tuberculosis and Polio trials: "Poster children" of medical RCTs

Grover (Russ) Whitehurst, Assistant Secretary for Research and Improvement at the U.S. Department of Education at the time, testified in front of the House Committee during the reapproval of the federal arm for education research:

> "If you look at medicine, for example, it's really only been within the last 75 years that medicine has become an evidence-based field. […] It was really the development of biochemistry, the science of physiology, which allowed medicine to get to the point where it had been a basic understanding of disease. Then it was the bringing on board of clinical trial experiments in the field in 1948 which have skyrocketed now to the point that there are 10,000 of them. That allowed medicine to take basic science and determine how it actually worked. We can do that in education. We need to do it." (Whitehurst, testimony, February 28, 2002).

When Whitehurst cited the year "1948" as starting point of medical RCTs, he was indirectly referring to the Streptomycin trial on Tuberculosis, which was the first published RCT in medical history (Armitage, 1995; Chalmers, 2001, 1162). Furthermore,

Whitehurst mentioned the thousands of trials since then that had transformed medicine into an evidence-based field, and he suggested that education could follow in its footsteps to become equally evidence based.

The Poliomyelitis trials have also been cited as a poster child of a medical RCT. During an U.S. Department of Education conference, Stephen Raudenbush, a USDOE adviser, referred to the Salk Poliomyelitis trial as having helped produce a "sea change" in medicine:

> One of the questions that comes up that's interesting is what caused the sea change in medicine and is it likely that anything like that might happen in education. That's way too big of a question for me to try to answer, but there is an interesting vignette, I guess, a part of the story that has to do with the Salk vaccine for polio.[…] But the results showed definitively that the vaccine was far more effective than not having the vaccine which led to further perfection, further clinical trials and ultimately the wiping out of polio as a disease. Now, we may not expect quite such dramatic success in saving lives in education, although the relationship between education and health is actually a very durable and interesting one, so maybe not being educated can cause a loss of lives. […] We need to learn how to do this. People didn't think you could do it in medicine. Like I said, the Salk vaccine trial was incredible, the double blind experiment. We need to be able to make the argument and we need to learn how to do this stuff. (Raudenbush, U.S. Department of Education Working Group, February 2, 2002).

The threat of Poliomyelitis was ingrained in the older generation's memories, as Raudenbush expressed: "Your parents would stand by in mortal fear as the doctor exercised your legs and did various things to see whether it was Polio" (Raudenbush, U.S. Department of Education Working Group, February 2, 2002). Thus, Raudenbush and others regarded the 1954 Poliomyelitis trials in the United States as both a victory over a crippling illness and as a powerful validation of the new RCT approach. Raudenbush pointed to the parallels between medicine and education: both save lives, metaphorically for education and literally for medicine, and the idea that either might experience a sea change was met with strong skepticism in both fields. He implied that a sea change could also happen in the field of education if it embraced the RCT technology as medicine did.

A closer look at the Tuberculosis and the Poliomyelitis trials reveals the historical role of these RCTs, and it demystifies their status as poster children. These trials are not perfect and positive models; they exhibited many challenges, which RCT promoters in education and international development rarely refer to.

Austin B. Hill's pragmatic perspective on the RCT model in medicine

Austin Bradford Hill was the medical statistician for the Streptomycin trial on Tuberculosis and was a major force in popularizing the randomized trial in the medical profession in the 1940s and 1950s. Although RCT advocates in education and international development do not necessarily quote Hill's works, they implicitly refer to his ideas when quoting the Tuberculosis trial.

Before Hill participated in the Tuberculosis trial, he had written the "Principles of Medical Statistics" (1937) for contemporary clinicians and social workers who had little mathematical training. In fact, Hill strove to make "obscure and repellent" statistics understandable in an elementary way (Hill, 1937, 2). In the foreword, the editor pointed out the necessity for understanding statistical principles due to the "growing demand for adequate proof of the efficacy of this or that form of treatment" (Hill, 1937, iii). That is, any assessment of success should be based on fact rather than opinion.

As a starting point, Hill addressed the distinction between the work of a laboratory worker and of a clinical researcher. In the clinical setting, the researcher could not control the many factors and multiple causes that influenced treatment effects (Hill, 1937, 3). Hill used the example of children who were in contact with measles, only some of whom had received a serum injection. Possible influences such as age, sex, social class, body weight, and state of health would need to be taken into account to determine whether the treatment prevented the illness. The researcher would need to either physically or statistically equalize the groups in every possibly influential or relevant respect, except for the serum treatment. The key problem was that no statistician would be aware of all the relevant factors.

> "If we find that Group A differs from Group B in some characteristic, say, its mortality-rate, can we be certain that that difference is due to the fact that Group A was inoculated and Group B was uninoculated? Are we certain that Group A

> does not differ from Group B in some other character [sic!] relevant to the issues as well as in the presence or absence of inoculation?" (Hill, 1937, 5)

Hill concluded that one can never be certain of not having overlooked relevant factors due to a "complex chain of causation" (ibid., 5). Hill then suggested "random allotment" of patients to treatment and control groups, so as not to introduce conscious or unconscious bias and to equalize the distribution of all characteristics. Hill had to make a general case for both concurrent controls and for randomization in particular (Armitage, 2003, 926). The practice of making fair treatment comparisons in medicine was not widely implemented and accepted by medical professionals (Chalmers, 2001, 1157); nor were the theoretical foundations established. Later, Hill recalled that he purposefully avoided the terminology of randomization, "because I was trying to persuade the doctors to come into controlled trials in the very simplest form and I might have scared them off" (Hill, 1990, 77). Although Hill tried to ease doctors' fear of statistical concepts, he anticipated resistance of medical professionals to controlled experiments.

Hill advocated randomization for pragmatic reasons due to selection bias: "Any deliberate choice of individuals to be treated may lead, unconsciously, to the treated group differing from the untreated group in some characteristic which, known or unknown, has an influence upon the results" (Hill, 1937, 8). He argued for concealed allocation schedules based on random numbers so that recruiters would not be able to influence assignment of patients to treatment and control groups (Chalmers, 2001, 1156).

As a pragmatist, Hill's writing was concerned less with the theoretical foundation of RCTs, and more with difficulties in clinical practice and experimentation. For example, Hill addressed ethical issues such as whether treatment could be justifiably withheld from patients (Hill, 1937, 7). He specifically pointed out that one cannot treat human beings like laboratory animals. To withhold from a patient a treatment that is likely to benefit him is morally wrong (ibid., 173).These issues are also important for the fields of education and international development. Hill argued that any new therapeutic measure should be given a trial period before coming into general use. Hill hoped that RCTs would effectively shorten the period of evidence generation and would prove to be time- and cost-effective by abbreviating the historical process of unsystematic observations

(ibid., 173). Hill cautioned that the new measure typically would have a relatively small effect that, even if important, might go undetected with small-scale tests. Therefore, a well-planned and extensive trial would be necessary to introduce new therapies. One issue for the researcher was determining when it was acceptable to withhold treatment to a patient for whom such treatment could *possibly* be beneficial, versus *probably* be beneficial (Hill, 1937, 173). Determining how much evidence is needed before initiating an RCT continues to pose serious challenges to researchers, both in the medical and social sciences.

Hill was also concerned with the generalizability of results. If one wished to argue from a sample to the "general run of patients," one would need to carefully consider whether the sample was fully representative of all patients, and not in any way biased or selected (Hill, 1937, 9). For example, if treatment was restricted to children with measles who managed to be in hospitals, results would not be representative of the "general run" of children who tended to be less ill and maybe of a lower social class (ibid., 11). Volunteers or self-selected individuals would not be a random sample of the general run of patients either. Hill thus foreshadowed many issues that would become important during the regulatory institutionalization of RCTs in the drug approval process; these issues would also confront RCTs in education and international development.

The Streptomycin trial on Tuberculosis as milestone in promoting RCTs

The Streptomycin trial was funded by the British Medical Research Council (MRC), which was the major public institution for medical research in the United Kingdom at that time. The council had been established in the context of the National Insurance Act of 1911 to make public research funding available and to "place in our hands new and more effective means of combating [these] diseases" (British Medical Journal, 1913, 1382). Medical researchers, epidemiologists, and statisticians teamed up to investigate new therapies, and ultimately they were able to convince their professional colleagues of the possibility for a clinical science (Kaptchuk & Kerr, 2004, 247; Armitage, 1995, 150)). A.B. Hill was one of these medical statisticians, and he was able to put randomization into practice with several medical experiments in the 1940s and 1950s, of which the

Streptomycin trial is the most well known. In his memoirs, Hill referred to the period as a "new era of medicine" (Hill, 1990, 78).

The investigators of the Streptomycin experiment called it the "first controlled investigation of its kind to be reported" (Medical Research Council, 1948, 780).[6] The trial worked toward a cure of pulmonary tuberculosis with the newly discovered antibioticum streptomycin, discovered in 1943. However, despite some clinical evidence, the MRC report found the evidence of effectiveness still "inconclusive" (Medical Research Council, 1948, 769). A trial with concurrent controls would therefore be justified. Hill would later justify randomization on different grounds—namely, the limited supply of the U.S. produced streptomycin in post-war England (Hill, 1990, 78). This justification would not have been necessary by true clinical equipoise, where the treatment outcomes would be unknown. However, the shortfall argument seemed necessary to convince physicians in charge of the Streptomycin trial to implement a control group without treatment, which gives an indication of the resistance to control group designs in the medical profession.

The research question was categorical: "Is streptomycin of value in the treatment of pulmonary tuberculosis?" (Medical Research Council, 1948, 780). The patients were randomly assigned for bed rest with streptomycin administration or bed rest alone, which would have been the normal treatment for this type of tuberculosis. Hill provided the randomization scheme. The trial was not blinded for doctors or the treatment patients. The control patients were not informed about the trial, an ethically questionable decision by today's standards.

For the Streptomycin experiment to be successful, an extended infrastructure was necessary. To gain access to sufficient patients, several hospitals and their physicians had to be recruited. For each site, a trial coordinator was trained on how to determine whether patients fit the research scheme based on certain inclusion and exclusion criteria. For example, the trial restricted the patients' age and type of tuberculosis (i.e., acute, fast

progressing, bilateral, not suitable for other therapies). Between January to September 1947, 107 patients were recruited. For six months, the streptomycin patients received four daily injections, whereas the control patients were only prescribed bed rest. The trial itself lasted for 15 months. The research team systematically collected and reviewed data from regular examinations and observations of toxic reactions.

The findings of the streptomycin study were positive. The measure of whether the patient died within the six months yielded a significant difference of 7% dead in the treatment group versus 27% dead in the control group. Based on more qualitative measures, 51% of the streptomycin patients improved considerably, in contrast to only 8% of the control patients. Both of these outcome measures were statistically significant at the .01 level. The report concluded that "streptomycin was the agent responsible for this result" (Medical Research Council, 1948, 780). More specifically, "streptomycin therapy was effecting [sic] what the patient's tissues alone could not do—checking the spread of the tubercle bacillus" (780).

Streptomycin was neither found to be a miracle drug nor did its absence prevent improvement. A control group was therefore warranted. Although the average treatment patient was better off, Streptomycin did not fully produce "clinical cures," and most patients still had the bacillus in their body. Deaths and radiological deteriorations happened especially toward the end of the trial. The authors attempted to explain why this could be the case. The infections, for instance, might have been too advanced in the first place. However, the authors were not able to systematically investigate this theory. Conversely, several control patients naturally improved their symptoms with combined bed rest due to the "natural recuperative power" (Medical Research Council, 1948, 781).

Challenges of the Streptomycin trial

I delineate several challenges of the trial from the report (Medical Research Council, 1948, 780): First, no answers about optimal dosage or duration could be given. Second, the selection of a homogeneous group of patients led to narrow findings. Third, trial

6 It is to be noted that although the MRC's 1944 Patulin trial had already used a randomization scheme, its results were published after the Streptomycin trial's article, probably due to its negative findings (Kaptchuk & Kerr, 2004).

procedures were changed midway, but its effect was unclear. Fourth, negative effects were not adequately captured.

First, although the trial could positively answer the general question of whether streptomycin was effective in treating pulmonary tuberculosis in a certain population, the trial could not answer questions about optimal dosage, duration, or degree of effectiveness for different variations of the illness. The report acknowledged that much additional research would be required "to determine the precise indications of streptomycin and the best schemes of dosage in pulmonary tuberculosis" (ibid., 781).

Second, the authors acknowledged that this one clinical trial was insufficient to fully determine the effect of streptomycin in different populations. They purposefully eliminated "as many of the obvious variations as possible" (ibid., 770) and instead followed closely defined criteria: "acute progressive bilateral pulmonary tuberculosis of presumably recent origin, bacteriologically proven, unsuitable for collapse therapy, age group 15 to 25 (later extended to 30)" (ibid., 770). The trial excluded, for instance, older and chronically ill patients. Selecting a homogenous patient group was based on expecting smaller variations in outcomes, which had the advantage of reducing the sample size and creating more precise results. The disadvantage was less generalizable results beyond the clearly defined group.

Third, the researchers made several adjustments to the Streptomycin trial in progress. Some control patients became eligible for a different treatment (i.e., collapse therapy) based on the course of their illness. The duration of the treatment was shortened from six to four months based on information from other clinics in the United States. The age requirements were changed. In general, adjustments allowed ethical treatment of patients (e.g., not withholding available alternative treatment when indicated), integration of new insights from other trials, and adequate recruitment numbers. These changes, though, may have affected the trial findings.

Fourth, negative effects of streptomycin presented further challenges. Toxic reactions were observed in most patients, especially physical coordination and vision (Medical

Research Council, 1948, 781). Systematic testing these reactions would have required additional trials. Weighing and balancing the positive and negative effects of the drug was an issue early on in medical trials.

Another challenge was streptomycin resistance, which the majority of treated patients developed. The report suggested adding another drug to decrease resistance (Medical Research Council, 1948, 781). Follow-up trials tested a combination of those drugs and found decreased resistance. The RCT approach allowed for efficiently introducing effective tuberculosis treatment and finding better dosages and drug combinations to avoid resistance in a fairly short period of time. However, clinical practice was slow to adopt the RCT findings (Cochrane, 1971, 80). In the 1970s, Archibald Cochrane complained that the medical profession continued to enjoy considerable freedom in treatment decisions without taking scientific findings into account (Cochrane, 1971, 82).

According to Valier and Timmermann, the success of the Streptomycin trial elevated the RCT to international prominence (Valier & Timmermann, 2008, 493). They identified key conditions as contributing to the rise of RCT, including: availability of funding; a well-developed infrastructure; and new organizational techniques that utilized interdisciplinary specialists and allowed for central data collection and review across multiple trial sites. All of these conditions held in the Streptomycin trial. The MRC's role was key in promoting and organizing cooperative trials. The question remained as to whether this new RCT technology would gain traction beyond the small group of medical academicians commissioned by the government (Meldrum, 2000, 1234). The MRC work, published in the widely disseminated *British Journal of Medicine*, soon gained a following in the United States at Harvard, Cornell, the National Institutes of Health, and the Veterans' Administration, among other places (Kaptchuk & Kerr, 2004, 250). Indeed, the new randomized approach eventually unified medical research (Valier & Timmermann, 2008, 494).

The popularization of the RCT in the U.S. Poliomyelitis trials

The Polio trials were among the largest and most publicized RCTs ever undertaken, and they demonstrated the "superior credibility" of the RCT approach inside and outside the

medical field (Meldrum, 2000, 1234). As mentioned above, the education researcher Raudenbush referred to the Salk trials as having helped produce a "sea change" in the medical science (Raudenbush, U.S. Department of Education Working Group, February 2, 2002).

A crucial fact has often been overlooked by researchers from other fields: Initially, the original Polio trial's design included only observed, non-randomized controls (Meldrum, 1998, 1234). In fact, Jonas Salk himself, the developer of the vaccine, did not advocate an RCT from an ethical point of view. Convinced that the vaccine would be effective and safe, Salk thought the trial was unnecessary. Denying the children the vaccine would have violated the Hippocratic Oath (Oshinsky, 2005, 180). The public health departments of thirty-six states also promoted observed controls (Meldrum, 1998, 1235). Statisticians and virologists, however, were against the observed control approach due to possible selection bias. Children from high-income and well-educated families were known to be more susceptible to Polio infection, but they were also more likely to volunteer for such a trial (Oshinsky, 2005, 177). Ultimately, the researchers reached a compromise and conducted two trials with two separate protocols—one using randomized assignment and one using non-randomized, observed controls. The RCT included approximately 400,000 children, and the one using non-randomized controls included approximately 950,000 children (Brownlee, 1955). To avoid possible public censure, researchers intentionally avoided the term "experiment" (Oshinsky, 2005, 191).

A year later, the results found the Salk vaccine to be 71 percent effective for the prevention of Poliomyelitis, with 60 to 90 percent variation depending on the strain (British Medical Journal, 1955). The larger, observational trial yielded similar results. Despite the results being positive overall, they were nonetheless disappointing. Ultimately, Albert Sabin replaced the Salk vaccine with a more effective vaccine a few years later (Oshinsky, 2005, 261). According to Oshinsky, pharmaceutical companies in the United States found the Sabin vaccine more profitable, and they successfully lobbied for its common use in the 1960s (Oshinsky, 2005, 325). In 2000, however, the United States passed a policy to only allow distribution of the Salk vaccine due to safety concerns (Hecht, Babcock, & Heymann, 2009, 59). Although scientific evidence played a

role in the history of Polio vaccines' implementation, other factors were important, including commercial interest, public sector planning, and scientific convictions..

Challenges of the Salk Poliomyelitis trials

The challenges of the Poliomyelitis trials were multifold: First, the prevention trial required a large machinery of operation. Second, researchers felt that the RCT was introduced prematurely and could only indirectly measure its effectiveness. Third, replication of findings proved difficult.

First, the struggle to commit to an RCT reflects the controversial status of random assignment among medical researchers. Much of this controversy stemmed from the sheer difficulty of conducting RCTs, given its large size and high costs. The Polio trial cost approximately five million dollars (unadjusted) (British Medical Journal, 1955, 1006). The large size of the RCT required a vast machinery of operational planning and execution, with 312 State and local health officials, 20,000 physicians and 40,000 nurses (Oshinsky, 2005, 189). Furthermore, the prevention RCT required a large population because the likelihood of contracting polio was only one in 2,000. Because children who were already vaccinated clearly no longer qualified as research subjects for additional trials, the possibility of holding additional Polio trials in the United States was reduced. This was one reason why Albert Sabin had to choose other countries such as the Soviet Union to test his vaccine. In fact, the Salk trial pointed to a general problem for human RCTs: the availability of research subjects for conducting RCTs. The rarer the condition being studied, the more difficult it is to secure sufficient numbers of research subjects.

Second, another important twist was that many Polio researchers felt that the trial was introduced prematurely. Salk had taken a controversial approach to vaccination. The findings from initial small experiments in disabled children had not yet been published or reviewed by outside experts (Oshinsky, 2005, 182). His main outcome was the rise in Poliomyelitis antibodies rather than permanent immunity, which would have been the only meaningful measure (Meldrum, 1998, 1234). Permanent immunity, however, could technically only be measured by purposefully exposing vaccinated children to the Poliomyelitis virus, an ethically questionable approach. At that point of research,

virologists found the safety risk to be too high, whereas the effectiveness of the vaccine remained questionable. The issue of the timing of a randomized trial has had a long history, of which the Salk trials are an illustrative case. Those safety concerns led to the British Medical Council deciding against an experiment on Poliomyelitis (British Medical Journal, 1955).

Third, upon review, the Salk vaccine was found to be safe and effective. In the subsequent year, however, the vaccine caused polio infections (Brownlee, 1955, 1005). Apparently, the testing and safety measures used in the production of the vaccination were relaxed compared to the measures used in the tested vaccine; they allowed virulent strains to enter the vaccine. This case illustrates the difference between experimental contexts and non-experimental contexts, the latter of which would have equaled distributing the vaccine to the general public. The experimental results were of relevance only insofar as the same product would have been produced and applied in non-experimental settings. Since the two products differed, the experimental findings were proven less relevant, ultimately leading to the replacement of the Salk vaccine.

Lastly, the Poliomyelitis trials used public schools as trial sites. Schools proved to be a logistically efficient way to recruit volunteers for the experiment. By utilizing public schools, the Polio trial became a "public experiment," garnering a large audience in the general public. The desperate need for a Poliomyelitis vaccine and the vaccine's ultimately positive results helped foster the public's acceptance of randomized trials as an evaluation tool. This also led education researcher Raudenbush to remember the Polio trials fondly and to promote RCTs in education (cf., Raudenbush 2002). Although formal RCTs had been practiced in medicine since the 1940s, RCTs were not required for the approval of new drugs until 1970.

2. The legal institutionalization of the RCT by the FDA

Thus far, medicine is the only discipline where RCTs have been publicly institutionalized and are legally mandated for evaluation purposes. Since 1970, regulations by the Food and Drug Administration (FDA) have required proof of effectiveness through two randomized trials before any drug is approved for public use.

The following four citations illustrate how education officials and researchers sometimes viewed the powerful role of the FDA as a model to which educational science should aspire. They saw the RCT as a key ingredient in making education policy as effective as the drug approval process.

An early example is the 1999 Brookings Institution's conference "Can We Make Education Policy on the Basis of Evidence?" The conference chair Paul Peterson from Harvard University made an interesting observation about federal education policy:

> We don't have anything like FDA. In medicine, FDA says you don't get approval unless you have survived the gold standard [the RCT]. We have yet to have any agency of the federal government, whether it is Congress or the executive branch agency or in any of our state governments, say in the field of education, yes, before you innovate, you've got to show that you've got an effective program here that deserves implementation on a wide scale. So FDA is the powerful instrument by which the concept of randomized experiment has shaped our whole understanding of what is the appropriate way of evaluating innovative procedures in medicine (Peterson, Brookings, presentation, December 8, 1999).

Peterson characterized the Food and Drug Administration as the "powerful instrument" that regulates drug approval via the RCT as the gold standard. Peterson envisioned a similarly powerful function for the U.S. Department of Education, but possible only if it had a similar gold standard for determining an education program's effectiveness. As the discussion later showed, Peterson saw the RCT as an equally suitable standard for education policy.

A few months later, during testimonies in the House Committee on Education and Workforce, Representative Michael Castle asked Reid Lyon about the possibility of restructuring of the U.S. Department of Education's research arm. Lyon had been in charge of the National Reading Panel's work on identifying effective reading programs (cf., chapter 3):

> Michael Castle: My question to you is because you have done hard-core scientific, medical research—can those same standards—the ones you look to in dealing with the FDA and various other hoops you have to go through to get medical research approved and then into usage be applied to education, or is that an overreach in terms of what we are doing?

> G. Reid Lyon: Absolutely, they can be applied.
> (Lyon, testimony, May 4, 2000)

There was a general optimism that the research arm of USDOE would be able to implement rigorous evaluation standards in education similar to what the FDA did for drug evaluations in 1970.

The "No Child Left Behind" law of 2002 and its push for scientifically based education raised hopes in certain groups that the national Institute of Education Sciences could become an equivalent of the FDA for education. In 2002, at a policy forum with the Education Secretary Rod Paige, Jon Baron's opening remarks likened the newest developments at federal education to the history of FDA.

> So, is No Child Left Behind the 2002 equivalent of the 1962 Food and Drug Administration amendments? We believe that the education policy community today is in a position that is similar to that of the medical community 40 years ago. You have a new law which says that funded activities shall be backed by scientifically based research including the preference for randomized trials. We believe that federal government's effective implementation of that concept has the potential to transform a field that has seen almost no progress in 30 years and create a new dynamic for evidence driven progress (Baron, presentation, November 8, 2002).

Baron interpreted the newest developments in education policy as important milestones, but lagging 40 years behind medicine. In 1962, FDA had demanded drugs to be effective, but it had not yet spelled out the standards of proof. It took another eight years for the RCT to become this standard in medicine. Baron implied that NCLB had not yet reached the breakthrough to transform education into a modern science as medicine had in 1970; but Baron was hopeful that the RCT would equally propel the field of education forward.

Lastly, in an interview, Grover (Russ) Whitehurst, the Director of USDOE's Institute of Educational Sciences, modeled the design of the What Works Clearinghouse after the FDA. It would be a government-sponsored registry for effective education programs:

> If you look at the U.S. Food and Drug Administration as a model, what's required to get FDA approval to market a product are two randomized trials. So, we will privilege randomized trials. We will provide a registry of those trials as related to particular Ps [products, practices, policies or programs]." (Whitehurst, T.H.E Journal, January 1, 2004)

These quotations in four contexts illustrate how much education officials and stakeholders desired the U.S. Department of Education to mimic the FDA and adopt its RCT standard. However, there is the question of to what degree discrete drug compounds and multilayered education programs could be evaluated with the same standards of evidence. Furthermore, even if they could, the FDA of 1970, which served as the reference point, was different from the FDA today. As shown in the final part of this chapter, the FDA Commissioner has been seeking to broaden the concept of medical standards—the standards which education officials try to emulate. RCT advocates in education seem to lag behind these newest developments.

The quest for "substantial evidence" in drug approvals

The Drug Efficacy Amendment of 1962[7] was an important step in the RCT history, because it strengthened the government's regulation for approving new drugs. The FDA, rather than the manufacturers, now decided when a new drug was introduced into the market. Notable physicians had testified that doctors in clinical practice could not evaluate the efficacy of drugs and that doctors often relied on a collection of impressions (Temin, 1980, 122). As a criterion for approval, the new law required "substantial evidence that the drug will have the effect it purports" before pharmaceutical companies could market the drug (PL 87-781, sec 5005d). "Substantial evidence" of effectiveness was legally defined as:

> Adequate and well-controlled investigations, including clinical investigations, by experts qualified by scientific training and experience to evaluate the effectiveness of the drug involved, on the basis of which it could fairly and responsibly be concluded by such experts that the drug will have the effect it purports or is represented to have under the conditions of use prescribed, recommended, or suggested in the labeling or proposed labeling thereof (sec 505d).

On one the one hand, this definition was expert driven, insofar as experts would decide what kind of evidence was needed. The FDA no longer saw the market as protector of its consumers, and it delegated the authority to choose among drugs from doctors to the government's medical experts.

7 Public Law 87-781, October 10, 1962. Amendment to the Federal Food, Drug and Cosmetic Act of 1938.

On the other hand, the definition of "substantial evidence" did not require agreement among experts for approval of a new drug. It was also unclear what should be concluded when a "weighty body of inconclusive or negative evidence" existed (National Research Council, 1969, 8). The definition of "substantial evidence" included the term "well-controlled investigation." Yet there was no agreement as to what constituted a well-controlled investigation (National Research Council, 1969, 8). It was not further defined until 1970, when two RCTs would be required for drug approval. Although the new standard still had unclear concepts, it was still more rigorous than the rule established by the Supreme Court's 1910 judgment, which had held that the basis of therapeutic effectiveness would be a matter of opinion (Temin, 1980, 125).

The FDA Commissioner decided that the Drug Efficacy Amendment of 1962 should be applied retroactively to all drugs approved between 1938 and 1962. Approximately 4,000 drugs with 300 distinct chemical formulas were on the market (National Research Council, 1969, 1). This resulted in the Drug Efficacy Study of 1969 by the National Research Council of the National Academy of Sciences. They assigned to each drug's therapeutic claim a categorical rating about the substantial evidence of effectiveness (ibid., 6).

The panel used the rating "effective, but …" to indicate the inferiority of a drug, despite its effectiveness, when they found drugs to be less effective than other drugs for a given indication (ibid., 9). Clinical investigations infrequently compared the effectiveness of a new drug with an old drug. Relative effectiveness between therapeutic agents thus went beyond the assigned task of the panel—an issue to be addressed in the Comparative Effectiveness Research movement (cf., end of chapter). The relation between safety and effectiveness presented another challenge. The panel did not offer firm guidance at that point, because the acceptable balance between benefit and risk would vary greatly with the use of the drug (National Research Council, 1969, 44).

The panel could not find well-controlled studies for many drugs despite the fact that these were widely accepted in medical practice. The pharmaceutical manufacturers only provided uncontrolled observations and testimonial-type endorsements as support for

their effectiveness claims. The panel had to decide how much weight they should give the "opinion of the marketplace" (ibid., 9). The report stated: "The final arbiter of the value of a drug is the consensus of the experience of critical physicians in its use in the practice of medicine over a period of years" (ibid., 9). The observational approach rather than the experimental approach still dominated medical practice.

The results of the investigation were that the panel rated seven percent of drugs as ineffective. Any drug not rated effective had to provide additional evidence to be continued. Some pharmaceutical manufacturers who had to discontinue their drugs went to court against the FDA. For example, the manufacturer of Panalba, a combined antibiotic, filed a court claim because the FDA had withdrawn its approval despite commercial success (Temin, 1980, 134). In this context, the FDA issued additional regulations defining the term "substantial evidence" (Food and Drug Administration, 1970). The FDA no longer advocated an expert model, but a clinical procedure model. Evidence could not be based on clinical observations alone. Historical controls were only allowed for diseases with high and predictable mortality (ibid., 7252). Instead, evidence had to be based on clinical trials with a treatment and control group combined with a systematic selection process. The method of selecting subjects "assigns the subjects to test groups in such a way as to minimize bias" and "assures comparability in test and control groups of pertinent variables." Steps had to be taken to "minimize bias on the part of the subject and observer" (ibid., 7251). Although random assignment was not directly mentioned in the regulation, it was implicit via minimizing bias. The United States government decided on the RCT approach to establish regulatory law for drug approval based on scientific rather than political authority.

The process from initial RCTs in medicine to the formal legal incorporation took several decades. The RCT mandate now requires the manufacturers to create an infrastructure that allows those trials. Although the U.K. Medical Research Council was publicly funded and had direct access to academic scientists, private sector manufacturers had no such infrastructure in place. The National Research Council attributed the lack of well-controlled studies to the industry's difficulty in commanding the needed clinical facilities and the services of experienced investigators in the United States (National Research

Council, 1969, 13). They felt that more national support in therapeutic research should be given, both in the programming and the management of trials (ibid., 13).

Since the 1970s, an RCT industry has developed in the United States in order to meet the changed regulatory requirements (Meldrum, 2000). From the 1980s onwards, medical research has relied on an almost exclusive use of RCTs (Brody, Miller, & Bogdan-Lovis, 2005, 517). Any U.S.-led clinical trial must be recorded publicly on the U.S. government's trial website. ClinicalTrials.gov recorded over 100,000 medical trials as of 2011.

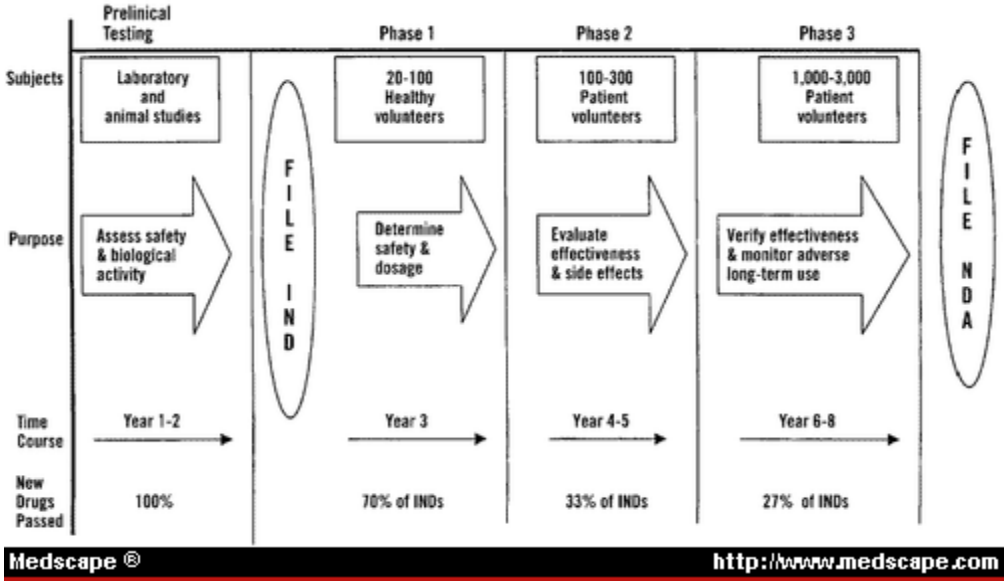
Current drug regulations requiring RCTs

When applying for FDA approval to market a new drug, companies will find that the current regulations explicitly refer to randomization: "Ordinarily, in a concurrently controlled study, assignment is by randomization, with or without stratification" (Code of Federal Regulations 21, 314.126). The FDA requires two positive RCTs with "substantial evidence" for drug approval.

Currently, clinical trials of new drugs are commonly classified into preclinical testing and four phases of clinical research. For approximately every 5,000 to 10,000 compounds that enter preclinical testing, only one is approved for marketing (Klees & Joines, 1997). The successful completion of RCTs of phase three is required for a New Drug Application (NDA) at the FDA (cf., Figure 1).

Before pharmaceutical companies can legally conduct clinical trials in humans, they need to submit an Investigational New Drug (IND) application to the FDA (Code of Federal Regulations 21 312.20, 4-1-2003 edition). The IND must include results from animal studies measuring preliminary toxicity, pharmacodynamic activity (i.e., what the drug does to the animal's body), and pharmacokinetic activity (i.e., what the body does to the drug).

FIGURE 1: Phases of drug evaluation



Phase 1 trials typically include 20 to 80 healthy volunteers to assess the safety and tolerability of the drug. Phase 2 trials investigate safety in a selected group of up to 300 patients. Phase 3 trials are what we consider the classical drug trials, involving cooperative multi-center studies with up to 5,000 subjects. Their goal is to determine the drug's efficacy, i.e., its effectiveness in the tested population. Phase 4 trials belong to post-marketing safety surveillance, and they address drug effectiveness, i.e., beyond the original test population—what A.B. Hill would call the "general run of patients" (cf., above).

One should note that there is an important distinction between the concept of "efficacy" and "effectiveness." Before 1970s, the terms "efficacy" and "effectiveness" had been often used interchangeably (cf., the Drug Efficacy Study). These two terms have acquired specific meanings in the current regulatory setting in the distinction of trial phases. Efficacy is the narrower term, referring to positive results within a restricted group of subjects often with homogeneous characteristics. Effectiveness refers to the concept of therapeutic success in a wider group of people.

The FDA requires "substantial evidence," not exclusive evidence. Two separate phase 3 trials need to show evidence of statistically significant positive results against a placebo treatment. Issues of effect size and comparator drug are secondary in the FDA approval process. The FDA focuses on internal validity; questions about generalizability and comparative effectiveness are secondary. These particular requirements influence how trials are designed in the first place. For example, pharmaceutical RCTs tend to be stronger regarding internal validity, the actual FDA focus, but less concerned with other possible biases.

Sources of bias in drug studies

Although in general RCTs minimize selection bias, they are not necessarily immune to other types of biases. Ernest House argued that despite the obvious advantages of RCTs, randomized drug trials are "among the most biased evaluations being conducted" (House, 2008, 416). Drug manufacturers would have an interest in producing positive findings because of their commercial interests. Trials sponsored by pharmaceutical companies and conducted by for-profit contractors were found to be three times more likely to have positive findings than trials sponsored by non-pharmaceutical, unaffiliated organizations (Lexchin, Bero, Djulbegovic, & Clark, 2003). In oncology studies in particular, 38 percent of studies funded by drug companies reached favorable conclusions, compared to five percent of studies from non-profit organizations (Friedberg, Saffran, Stinson, Nelson, & Bennett, 1999). This would lead to the conclusion that investigators have some control over the direction of the findings. House identified 14 sources of bias that would result from opportunistic choices in drug studies (cf., Table 2). The sources of bias do not pose threats to internal validity; thus no bias or systematic error in the narrower sense would exist. House used the term "bias" in a wider sense, applying it to threats to external validity and conflict of interest, whereas Howard Brody used the term "commercial bias" to describe such distortions (Brody et al., 2005, 578). This commercial bias can be found in opportunistic choices of comparison group, dosage, administration, sample, and time scale—all of which can affect external validity of findings. For instance, using a less effective dosage for a comparator drug would yield more effective findings for the experimental drug.

TABLE 2: Opportunistic choices in pharmaceutical RCTs

Opportunistic choices	Examples	Domain
Choice of comparison	Placebo rather than state-of-the-art drug as comparator	External validity
Choice of dosage	Less effective dosage for comparator drug	
Choice of administration	Less effective administration of comparator drugs	
Choice of sample	Homogenous sample; subjects with fewer comorbidities	
Choice of time scale	Chronic-use drugs tested for short periods	
Redefining outcomes after findings to achieve success	Retroactive change of primary outcome	Interpretive bias
Choices in data analyses	Providing researchers with incomplete data sets	
Favorable interpretations	"This drug is now the treatment of choice."	
Choice of outcomes	Choosing outcomes most favorable to hypothesis	
Underreporting of negative effects		Reporting bias
Underreporting of unfavorable data		
Control of authorship	Company employees, rather than researchers, writing reports	
Selective publishing	Publishing only positive findings	
Deceptive publishing	Publishing positive findings repeatedly under different authors	

Adapted from House, 2011, 70.

Drug testing—as any other evaluation activity—typically starts with a confirmation bias, i.e., the assumption that the drug to be tested would be effective. Clearly the sponsors would not invest the necessary resources if they believed the likelihood of effectiveness to be low. Confirmation bias may guide decisions involved in the planning, designing, and interpreting of an RCT. Even without financial incentives, individual researchers who are vested in their research trial are likely to have some confirmation bias. In the Polio testing, for instance, Jonas Salk did not regard an RCT necessary as he was already convinced of the drug's beneficial outcome (cf., above). Such preconceived expectations would then influence the particular outcomes and findings of a trial; i.e., what a researcher is looking for seems to affect the research outcome (cf., Kaptchuk, 2003, 1454). The internal validity of the result itself may not be affected, but the validity of the

experiment may be limited in other ways, e.g., with respect to generalizability and relevance to non-clinical settings.

Researchers have to make many decisions about the design of RCTs. As Fisher had pointed out, many considerations go into the design of an experiment beyond statistical expertise (cf., previous discussion). These considerations will influence the outcomes to some degree. For example, the choice of what a drug is compared to may influence findings. Placebo trials do not allow for comparative effectiveness research. From a comparative effectiveness perspective, head-to-head studies with the standard-of-treatment drug would be preferable. Pharmaceutical companies, however, like to use a placebo instead of the standard-of-treatment drug as comparator (House, 2008, 417). The likelihood of positive results is higher with a placebo, because the comparator situation is the natural progress of the disease rather than a drug already established to be effective. The choice of placebo as comparison treatment would only be ethically legitimate when no standard of treatment existed. This was the case with the Polio trial and the Streptomycin trial. If there were an alternative treatment, such as in the antipsychotic example discussed below, then a placebo group would not be ethical.

Even if the standard-of-treatment comparator drug is used, a dosage that is too high or too low for standard treatment may be disadvantageous. For example, a second-generation antipsychotic was tested using the standard treatment of Haloperidol as comparator. However, the dose was unreasonably high, leading to increased side effects and making the new drug look more tolerable (Carlat Psychiatry Report, 2009).

Another way to increase the chances of positive findings is by reducing sample heterogeneity and by choosing low-risk patients, the latter of which are more likely to respond positively to the drug. Thus study entry criteria may be very strict. For example, the clinical trials that led to the FDA approval of Seroquel for bipolar depression disorders excluded all patients who had comorbidities and who had been through at least two treatments of antidepressants; even though patients diagnosed with bipolar II who were seen in clinical practice were likely to exhibit such comorbidities (Thase, Macfadden, Weisler, Chang, Paulsson, Khan et al., 2006). It is therefore hard to conclude

that Seroquel would be effective for patients such as those excluded (Ashih, 2009, 1–2), because the clinical subjects were not representative of the target population seen on a daily basis.

Sources of bias may also enter in the publication phase. Positive findings are more likely to be published, whereas negative findings are more likely to be suppressed. Although companies must submit all clinical results to the FDA, the company can choose which results to publish. Selective publishing involves either under-reporting or over-reporting findings. Underreporting happened with the use of antidepressants among children. In 2004, a pharmaceutical company was accused of suppressing results of four trials that showed an increased risk of adverse effects among children using antidepressants (Lancet, 2004). Conversely, bias in over reporting took place with the antipsychotic Risperidone. One trial was reported in six different publications with different authors' names for each (Pearson, 2007). Furthermore, positive results for only certain subgroups, rather than the entire trial population, may be published in academic journals. These practices would not affect the FDA approval process, but it would influence medical practice, as medical professionals might assume an independent trial for each publication.

The point I wish to underscore is that although these choices and practices by evaluators do not influence the narrow concept of internal validity, they do influence the results of a trial and how they are perceived. Opportunistic choices in the design and the interpretation of RCTs may lead to more positive findings, but such choice may also decrease the relevance of the findings for—as A.B. Hill put it—the "general run of patients."

3. Evidence-Based Medicine and the RCT reference standard

In the 1990s, RCTs were further mainstreamed through the Evidence-Based Medicine (EBM) movement. Although the RCT was widely anchored in medical research as the preferred evaluation approach, medical practitioners still tended to rely on unsystematic observations based on clinical experience. The EBM group emphasized the RCT standard in medical practice and training. They recognized that they could not prove their own claim via an RCT of the superiority of evidence-based medicine over traditional

approaches (Evidence-Based Medicine Work Group, 1992, 2424). This fact illustrates limited applicability of RCTs in different evaluation domains. Evidence-based systematic reviews provided professional physicians with tools to make evidence-based choices. Howard Brody referred to a power shift within medicine, where the medical scientist guides the medical practitioner (Brody et al., 2005, 572). The legal institutionalization of the RCT was finally brought to the doctor's office.

Archibald Cochrane, a British physician who had studied medicine under the influence of Austin B. Hill, observed that Hill's ideas of randomization had only made small gains in medical practice. In his book "Efficiency and Effectiveness" (1971), Cochrane deplored the notion that clinical opinion, the oldest form of medical evidence, would hold more weight than an experiment. Cochrane's ideas in 1971 could be seen as a prototype for the evidence hierarchy that would come later, which was based on the RCT as reference standard. According to Cochrane, the double-blind RCT would solve the problem, in that patient characteristics would be randomized between the two groups; neither the doctor nor the patients would know which of the two treatments was given (Cochrane, 1971, 22–23). For Cochrane, clinical opinion would be the worst type of observational evidence; it lacked quantitative measurement or any attempt to discover what would have happened if the patient had received no treatment (Cochrane, 1971, 21). In such instances, opinion was followed by observations with no control group. Comparison groups were the next step, but those without randomization were often prone to selection bias. Cochrane referred to comparison groups as a "mixed lot" (Cochrane, 1971, 21). For example, people who refused treatment could end up in the control group. Therefore, an RCT was warranted.

Based on Cochrane's ideas, an international group from the McMaster University in Ontario, Canada, formed the Evidence-Based Medicine Working Group in 1992. The purpose of the group was to promote putting clinical medicine on firmer scientific footing. They proclaimed a "NEW paradigm for medical practice" in the *Journal of the*

American Medical Association—a journal widely read by medical clinicians (Evidence-Based Medicine Work Group, 1992).[8]

The group defined EBM as "the process of systematically finding, appraising and using contemporaneous research findings as the basis for clinical decisions" (Rosenberg & Donald, 1995, 1122). The group suggested a new way for medical practitioners to access medical literature. The idea was that a physician would systematically move from a problem statement about an individual patient to the process of critically appraising the literature to arrive at a relevant, statistically valid answer. This would ultimately result in superior patient care and better patient health (Evidence-Based Medicine Work Group, 2424). Medical residents would be trained to understand the methodological criteria to systematically evaluate the validity of the clinical evidence and to use quantitative techniques for summarizing the evidence (Evidence-Based Medicine Work Group, 1992, 2421). The 1992 article gave the ideal example of how this method would look in practice: After examining a patient, a resident went to the library searching for and retrieving articles about seizures; upon finding the answer, she then conveyed the answer to the seizure patient.

The EBM group did not regard the RCT as the only way to establish evidence, yet it identified the RCT as the reference standard. The group arranged levels of evidence in a hierarchy, in which the RCT constituted the top, optimal approach to evaluation. Different hierarchies of evidence have been developed including the U.K. Clinical Guidelines for Primary Care, illustrated in TABLE 3.

8 One might ask whether they truly introduced a paradigm in Kuhn's tradition, which would have required a shift in the conceptual world view that determined methods, research questions, problems considered relevant to solve, and new standards of evidence (Kuhn, 1962). They seemed rather to be applying medical research standards that already existed into the field of medical practice (Solomon, presentation, June 19, 2009).

TABLE 3: Hierarchy of medical evidence

Grade	Type of Evidence: Evidence from...
I a	Systematic review of RCTs
b	Individual RCT
II a	One controlled study without randomization
b	One other type of quasi-experimental study
III	Observational studies
IV	Expert committee reports or experts

Adapted from http://www.eguidelines.co.uk.

The grading of medical evidence is based on the methodological approach of an evaluation. Randomization is the distinctive difference between grade 1 and grade 2 evidence. The best evidence for a medical treatment would be a comparative review of RCTs. First-grade evidence relies on at least one RCT. Grade 3 and 4 evidence includes observational studies and expert opinions. Because these do not include a comparison group, they are considered biased and generally insufficient grounds on which to base clinical decisions. For many clinical questions, however, high-quality evidence is not available; such questions would need to consider weaker evidence (Evidence-Based Medicine Work Group, 1992, 2424). Other medical guidelines consider systematic reviews of RCTs also as the highest form of evidence. In 2011, the University of Oxford's Centre for Evidence-based Medicine[9] revised their hierarchy, in which they distinguished high-quality and low-quality RCTs. Low-quality RCTs belong to Level 2 evidence and fair worse than systematic reviews of cohort studies.

To put EBM into practice, the International Cochrane Collaboration (ICC) was founded in 1993.[10] Its goal was to provide medical practitioners with ready-made analyses of the effectiveness of health interventions. ICC has produced systematic reviews in the form of high-level overviews of primary research on a particular research question. To answer the research question, a systematic review attempts to identify, select, synthesize, and appraise the available RCT evidence. A review could generate aggregate evidence of

9 Oxford Centre for Evidence-based Medicine, http://ww.cebm.net, accessed June 14, 2011.
10 International Cochrane Collaboration, http://www.cochrane.org/, accessed June 13, 2011.

several small-scale studies that might not be statistically significant on their own. One drawback of the systematic review is that it can only rely on existing primary studies. It is not able to add more substantive information than that which is provided by the individual studies.

An example of an ICC review concerns the use of antibiotics for a sore throat. The review found 27 experimental studies with a total of 12,835 cases of sore throat that met the researchers' selection criteria. The review determined an average reduction of symptoms by 16 hours, but it also cautioned against the side effects of antibiotics (Spinks, Glasziou, & Del Mar, 2006). Therefore, absolute benefits of antibiotics use were modest. A similar review of antibiotics for acute laryngitis found merely two RCT studies with a total of 106 patients with few benefits (Reveiz, Cardona, & Ospina, 2007). Nonetheless, the prescription of antibiotics for these symptoms was discouraged, based on the limited RCT evidence.

ICC developed a protocol for how to extract and weigh evidence from different sources. The "Cochrane Handbook for Systematic Review" (revised in March 2011) laid out how to determine the validity of results. Risk of bias was regarded as a key concern. Bias as systematic error could appear in multiple ways: selection, performance, detection, attrition, and reporting (cf., TABLE 4).

TABLE 4: Classification scheme for bias

Type of bias	Description	Relevant domains
Selection bias	Systematic differences between baseline characteristics of the groups that are compared	Sequence generation, Allocation concealment
Performance bias	Systematic differences between groups in the care that is provided, or in exposure to factors other than the interventions of interest	Blinding of participants and personnel
Detection bias	Systematic differences between groups in how outcomes are determined	Blinding of outcome assessment
Attrition bias	Systematic differences between groups in withdrawals from a study	Incomplete outcome data
Reporting bias	Systematic differences between reported and unreported findings	Selective outcome reporting

Adapted from the Cochrane Handbook 2011.

TABLE 4 is meant as a tool for systematic reviewers to gauge the quality of an individual RCT, with a focus on internal validity. RCTs may still exhibit several biases despite using randomized assignment. First, selection bias occurs when systematic differences exist between baseline characteristics of the groups that are compared. Randomization of participants typically minimizes this bias; however, if randomization is not properly performed due to irregularities in the sequence generation or weak allocation concealment, selection bias may still prevail. Second, performance bias can happen when no blinding is used, which means that participants and personnel know who is in the treatment group and who is in the control group. However, blinding is often not possible for invasive treatments such as surgery. Blinding is hardly feasible for educational and international development interventions, and thus perception bias could result. Third, detection bias may arise from not blinding outcome assessment data, which could result in recorders observing outcomes more favorably for treatment subjects than for control subjects. Forth, attrition bias could occur due to the withdrawal of participants who are different from the average person in the treatment and control groups. Therefore, the original average similarity of characteristics between treatment and control subjects would not hold, thereby reducing the internal validity of findings. Finally, reporting bias arises from selectively reporting outcomes. Systematic differences between reported and unreported findings exist. These differences suggest that significant findings are more likely to be reported.

Note that TABLE 2 was concerned with biases beyond those regarding internal validity, including external validity. Recall that the evaluation theorist Ernest House argued that randomized drug trials had been among the "most biased evaluations" (House, 2008, 416). Commercial pharmaceutical trials are more likely to exhibit such biases to enhance positive findings and reporting. For example, RCT usage did not prevent drug sponsors from engaging in opportunistic choices that might favorably influence RCT findings.

Such a classification scheme of biases is useful for determining the evidence level of medical findings beyond just internal validity, which has been shown to be the primary indicator of efficacy in federally approved drugs. Such schemes would also be beneficial

when using RCT findings in the fields of education and social development. Due to the impossibility of blinding, for instance, performance bias is of great concern for any RCT in the evaluation of educational and social programs. The study subjects may change their perceptions and behaviors based on the fact of whether they were selected to be in the treatment or control group. Randomization at the community or school level may mitigate the problem of changing self-perception. Then, however, the sample size must be increased several times to obtain statistically significant results, based on fewer, randomly assigned units. The number of research units directly affects p-values and effect sizes. Therefore, careful considerations of the pilot size are required to obtain valid and policy-relevant results.

4. Relaxing the medical RCT model to influence future regulatory policy?

In the field of medicine, discussions about RCTs have encompassed both negative and positive reactions from academics and clinicians since their occurrence in the 1940s (Straus & McAlister, 2000, 387). The resistance to using RCTs for determining effectiveness, however, has been weaker in medicine than in other policy areas. Nevertheless, there has still been a sizable group within the medical profession that has been skeptical about the privileged status of RCTs.

The following section first reviews criticisms of the RCT model raised by ethicists, philosophers, and clinical professionals. It then features voices from regulatory agencies in the United Kingdom and the United States that have distanced themselves from the privileged status of RCTs and that argue for a more integrated approach to medical science. The section concludes by describing how the Comparative Effectiveness Research movement illustrates the practical need for a more comprehensive science in medicine to answer fundamental questions—especially regarding personalized medicine tailored to an individual patient's needs.

Critical responses to the RCT model in medicine from within

Even RCT-supporters caution against the uninformed use of RCTs. Archibald Cochrane, for example, praised the RCT as a "beautiful technique, of wide applicability, but as with

everything else there are snags" (Cochrane, 1971, 22). For example, by its very nature statistical significance may lead to wrong conclusions; a significance level of five percent would necessarily generate one misleading result out of twenty. In addition, large studies with large sample sizes may achieve statistically significant results with only small effects, which may be clinically unimportant (ibid., 23). Those considerations would be important in the critical appraisal process of the literature.

In particular, there has been much subjective judgment surrounding the selection of research subjects or surrounding the measurement categories of drug effectiveness. Table 2 summarized the different opportunistic choices in pharmaceutical RCTs, such as when choosing a comparison treatment or patient characteristics. The oncologist W. Hilbe highlighted the potential for optimizing results via inclusion and exclusion criteria for patients to participate in clinical trials (Hilbe, 2010, 1). It is known that narrowly targeted treatment groups typically improve the efficacy and precision of RCT results. In particular, younger subjects tend to participate in trials whereas the elderly are often excluded in RCTs (Simon, 2001, 940). The narrower the target group of an RCT, the less its results can be extrapolated to other individuals (Simon, 2001, 940). This practice of excluding certain populations could lead to a point where the tested drugs could only be applied to a minority of patients unless clinical physicians engaged in off-label treatments.

The occurrence of biological variation within the human species has hampered attempts to extrapolate evidence from the study population to other individual patients (Straus & McAlister, 2000, 388). Clinical uncertainty exists about applying the right treatment to the right patient at the right time (Conway & Clancy, 2009, 328). Which interventions are most effective for which patients under which circumstances? Conversely, how is an individual's response different from the average patient in a trial? RCTs are not the best tool to answer these questions of "particularization" because they deal with the "average patient" and they do not generate an understanding of the reasons for the differences in outcomes within the study group (Bluhm, 2005, 537). In order to make an informed decision about a particular patient, the medical practitioner would need additional knowledge of biological factors that might influence drug effectiveness. The "bare

bones" approach of experimental methodology does not suffice for delivering this knowledge (Will, 2007, 87). This type of information about underlying biological mechanisms, however, comes from biological research and observational studies, which EBM typically ranks lowest on its evidence hierarchy. Thus, uncovering the "causal stories linking exposures with health outcomes" might become an important contribution to experimental research (Bluhm, 2005, 543). Based on this line of thinking, the theorist Robyn Bluhm favored a combination of experimental and observational laboratory research for providing additional evidence to understand drug and intervention effects.

EBM, which has generally adhered to an experimental evidence hierarchy (cf., TABLE 3), faced critics from its beginnings in the 1990s. From a more ethical-institutional perspective, the medical ethicist Howard Brody and his coauthors reviewed arguments of "enemies" and "friends" of evidence-based medicine (Brody et al., 2005). Brody and coauthors identified one major criticism—i.e., that EBM had shifted the focus of medicine from the individual patient to a larger population (Brody et al., 2005, 570). Population research is primarily concerned with average results obtained from large groups of people. Although the unit of analysis is the individual, RCTs produce average effects for the average patient rather than for the individual patient. The estimation of a drug's effect in a study population does not necessarily translate to one individual. It is therefore often unclear whether an RCT finding can be applied to a particular individual patient.

Regarding RCTs and individual decision-making, Brody and colleagues outlined the common argument that the EBM model favoring RCTs does not offer the choice of treatment to the individual patient as research subject. In particular, the nature of randomization does not allow individuals to make a choice as to whether to be in the treatment or the control group. Few physicians want to randomly assign treatment to patients, to forgo the individualization of therapy, or to withhold promising new therapies from a particular patient (Kaptchuk, 1998, 1723).

From a profession's point of view, the increased focus on RCT evidence had denigrated clinical expertise and undermined the shared physician-patient decision-making process

(Brody et al., 2005, 574; Straus & McAlister, 2000, 389). Brody and colleagues argued that shifts in power and authority triggered certain criticisms of EBM. Medical clinicians had trusted first-hand clinical experience and had sometimes favored interventions without strong reliable evidence to support the interventions. They tended to overestimate the actual variations among individual cases when they made judgments based on individual clients (Brody et al., 2005, 571). The EBM model does not favor clinician's opinions and observations—a fact Cochrane had already stated in 1971.

Brody and colleagues observed that some EBM practitioners, sponsors, and bureaucrats uncritically accepted the RCT as a sole source of evidence, while rejecting all other forms of evidence (Brody et al., 2005, 570). This so-called "crude EBM model" elevated the RCT to a nearly sacrosanct status served as a straw man for critics (Brody et al., 2005, 573). This meant that supposed allies of EBM misrepresented the EBM model and opened it to criticism. Brody and colleagues stated correctly that EBM supporters might have simplified and misrepresented the true nature of EBM, and therefore, critics easily pointed out the deficiencies in the currently practiced EBM and its reliance on experimental evidence. However, Brody et al. do not sufficiently address shortcomings of EBM in their article. In particular, they left out the issue of RCTs' inability to address questions surrounding the causal mechanisms of the treatment process.

New developments in regulatory medicine: a shift away from the privileged RCT?
The U.K. National Institute for Health and Clinical Excellence (NICE) has been a major trend-setter in medical research and policy. Its chair, Sir Michael Rawlins, pointed out several limitations of randomized controlled trials (Cuthbertson, 2008). Rawlins argued that RCTs have limited external validity because they are typically used for "specific types of patients for a relatively short period of time." In medical practice, however, treatments would be needed for a wider variety of patients and for longer periods. For example, RCTs would often exclude comorbid patients, despite the fact that the average patient often has more than one illness. RCTs with patients having heterogeneous characteristics, however, were more likely to produce insignificant results. Restricting RCTs to a narrow set of patient characteristics would increase the likelihood of finding

statistically significant positive results. Drug sponsors would therefore have minimal incentive to choose more heterogeneous populations.

Rawlins stated that RCTs had been put on an "undeserved pedestal" (Cuthbertson, 2008). He questioned the status of RCTs at the "top of hierarchies." His metaphor of the pedestal refers to a vertical concept of evidence assessment, in which there would be a clear winner at the apex. As Rawlins argued, "hierarchies are illusionary tools for assessing evidence" (Cuthbertson, 2008). A danger was that "hierarchies would attempt to replace judgment with an over simplistic, pseudo-quantitative assessment of the quality of the available evidence." As a result, decision makers would adopt an "entrenched position about the nature of evidence," which would influence and dominate how medical decisions are made. Rawlins suggested that the concept of hierarchies should no longer be used as a heuristic tool. In a similar way, Bluhm had argued against an experimental hierarchy of evidence generation. She suggested different hierarchies for different types of clinical questions, and she argued that observational studies should be given more attention. Rather than lumping those into one sub-standard level, Bluhm argued for a more distinctive categorization of observational evidence, such as cohort studies and case-control studies (Bluhm, 2005, 536). Likewise, Rawlins recommended that RCT hierarchies "should be replaced by a diversity of approaches that involve analyzing the totality of the evidence base" (Cuthbertson, 2008). For example, observational studies including historical controlled trials and case-control studies could be important sources of evidence. Rawlins, who is from the U.K. Institute of Health, may well become a trendsetter for modifying the RCT-laden evidence hierarchy. Similar concerns can be heard in the United States.

Voices within the U.S. Food and Drug Administration had also questioned the current dominance of RCT in medical innovation in several speeches. Margaret Hamburg, the U.S. Commissioner of Food and Drugs, expressed an interest in a personalized medicine in which the right therapies are tailored to the right people (Hamburg, speech, February 25, 2010). Personalized medicine is about understanding that people differ in their genetic makeup, their environment, and their lifestyle; and these differences are critical factors in the individuals' diseases and how these individuals respond to therapies. For

example, genetic markers can indicate possible benefits of certain therapies, as in cancer patients.

In the context of personalized medicine, Hamburg rethought the level and kind of evidence needed for drug approval and design of drug trials (Hamburg, speech, February 25, 2010). She called for more flexible standards of product evaluation and optimization of clinical trial designs (Hamburg, speech, May 15, 2010, Hamburg, speech, October 13, 2010). For Hamburg, a science-based regulatory agency, such as the U.S. Food and Drug Administration, would need to rethink its currently inefficient drug approval procedure. Results of unsuccessful clinical randomized trials in the past might be revisited through the lens of new biomarkers. New diagnostics might turn such RCTs into positive findings for subpopulations of responders with certain genetic markers (Hamburg, speech, February 25, 2010). In a speech to the Organization of Economic Cooperation and Development, Hamburg remarked that "to strengthen the international clinical trial paradigm" would also mean to look "beyond the randomized controlled trial" to additional approaches for ascertaining safety and efficacy (Hamburg, speech, September 27, 2010). Hamburg quoted observational studies that rely on the use of a medical product in the course of a medical practice, as well as meta-analyses, post-marketing surveillance, and data-mining techniques (Hamburg, speech, September 27, 2010). Hamburg pointed out new clinical trial methodologies that could balance methodological rigor with the need for more rapid and more targeted answers. For example, Hamburg suggested studying smaller populations than those studied in traditional RCTs. This would be possible by using modeling and simulation techniques and by applying statistical methodologies and protocol designs for using real-world data from registries and healthcare databases (Hamburg, speech, October 13, 2010).

The call for observational studies and statistical modeling of real-world data reflects a shift within scientific thinking in medicine. As FDA Commissioner, Margaret Hamburg has been a frontrunner in what might become commonly accepted standards and knowledge in the future. This could go hand in hand with a devaluation of the RCT, or, expressed positively, in a reconsideration of other methodologies—including observational studies– that the RCT movement had dismissed in the past. New advances

in alternative methodological approaches might have contributed to the renewed interest of these methodologies. For example, new insights into genetic markers that directly influence patients' reactions to certain therapies might not need to rely on results from large-scale clinical trials.

There is a general question of how to produce knowledge in medicine. Policy makers and researchers in the United States consider Comparative Effectiveness Research (CER) as potentially being a more relevant approach to knowledge generation and dissemination in medicine than RCT-centered EBM research with zero-treatment control groups. The 2009 American Recovery and Reinvestment Act included in the U.S. economic stimulus supported CER and led to the Patient-Centered Outcomes Centered Research Act of 2009 (Conway & Clancy, 2009, 328, 330). CER is a variety of Evidence-Based Medicine (EBM), but with important differences. CER precludes an immediate adoption of an evidence hierarchy, and it does not necessarily privilege RCTs over all other methods. CER thus provides greater methodological flexibility and includes pragmatic trials in routine clinical practice, observational findings, and modeling. CER also focuses on subgroup analysis; RCT researchers, in contrast, avoid subgroup analysis in most cases, or they might only use it cautiously because the original random assignment did not take place according to those subgroups (Simon, 2001, 942). The goal of CER is to ensure external validity by focusing on characteristics of subgroup populations (Tanenbaum, 2009, 976, 977). CER also attempts to address the common exclusion of heterogeneous patients from RCTs (Tanenbaum, 2009, 977).

It is too early to judge whether the new movements of personalized medicine and comparative effectiveness research will have a lasting impact on the medical science and policy. In general, the renewed attention to the value and challenges of RCTs has been a productive process that addresses relevant policy questions relating to individual patient care.

Outlook: Learning from the medical field about RCT use

Many stakeholders in education and international development policy view the privileged status of the RCT model in medicine as a desirable situation for their own discipline, thus

hoping to gain more credibility and authority for their own actions. However, they tend to overlook the challenges of the RCT model even within medicine, though they could potentially benefit from understanding how the medical field addressed such challenges.

One challenge, for instance, is the appropriate timing of RCTs, thought they have had a long tradition in the medical sciences. If all chemical compounds were tested by an RCT, the medical scientists would rarely find positive results. Instead, they are tested in preclinical, then phase 1, and phase 2 trials, all of which often rely on non-experimental observations. Only then, the few promising chemical compounds will be tested in large-scale RCTs (phase 3). Medical RCTs are thus timed in later stages of the evaluation process. In education and international development, however, timing is often a neglected issue. This may result in premature RCTs, similar to what occurred in the Salk Poliomyelitis trial, where initial observational results were not available. Therefore, sufficient pretrial indications of a program's possible effectiveness are needed before an RCT could be deliberated for determining program effect.

A second challenge is evaluating negative side effects. Early on, medical RCTs have not only been concerned with effectiveness, but also with safety and side effects. The Streptomycin trial of 1948 did not yet systematically assess the negative side effects such as impaired physical coordination and vision. But since the 1962 Drug Efficacy Amendment, the FDA has been equally concerned with drug safety and drug efficacy. Evaluating negative effects of interventions, however, has not been at the forefront in education and international development.. Their underlying assumption is that interventions have either a positive or zero effect, but not a negative effect. This is a mistaken belief, especially when an intervention is compared head-to-head with another intervention, which is rarely done.

A third challenge, which occurs in invasive medical treatments such as surgery, is dealing with an unblinded study population. Knowledge about what group the patients belong to may affect their perception and spontaneity and thus influence trial results. Blinding is hardly feasible for educational and international development interventions. As I

illustrate in the next chapter, the Student Teacher Achievement Ratio (STAR) trial was not blinded, a fact which could have influenced the teachers' motivation to teach.

These three challenges are examples of how applying RCTs in medicine could also inform RCT evaluations in education and international development. In the next chapter, I analyze how the U.S. education system has attempted to make the RCT the privileged method for evaluating federal education programs. Certain challenges of medical RCTs apply even more to the educational field, where the tested population comprises students instead of patients, and the treatments are comprehensive programs instead of drugs.

CHAPTER 3: THE MEDICAL RCT MODEL IN EDUCATION POLICY

RCT advocates in education often point to important successes of clinical trials in medical research, such as the Streptomycin and Poliomyelitis trials (cf., chapter 2). However, their portrayals of these uses of RCTs typically gloss over challenges important to these methods even in their best cases. These challenges prove to be even more difficult in education. I show that, in particular, the limited representativeness and generalizability (cf., appendix) of RCT findings have posed major roadblocks to policy development in education.

In this chapter, I take a closer look at three areas where the RCT model influenced education research and policy: first, the Student Teacher Achievement Ratio experiment in Tennessee; second, the National Reading Panel's adaptation of the RCT-based medical model; and third, the Reading First initiative's struggle to implement RCT-backed policy. In all three cases, applying the RCT model proved to be difficult. Providing a more nuanced view of these cases may help policy makers discern appropriate lessons for education policy and inform their future quests for scientifically-based policy making.

I begin in the first section by considering the influence of the Student Teacher Achievement Ratio (STAR) experiment in Tennessee, which has been the poster child for the possibility of successful RCTs in education. As a pilot, STAR faced challenges in producing policy-relevant findings and in applying these findings to other contexts. The California Class Size Reduction Program illustrates how the extrapolation of findings from the STAR pilot failed in the new context of a statewide initiative. Despite these setbacks, policy makers and researchers have cited the STAR experience in the quest to champion rigorous RCT research in education policy.

In the second section, I shed light on the federally-mandated National Reading Panel (NRP) and its attempts to adapt the medical RCT model for evaluating reading research from 1998–2000. NRP's work inspired RCT-based meta-reviews in education and in the work of the federally funded What Works Clearinghouse (WWC) in order to guide local decision making in education. The direct influence of the WWC on educational practice,

however, proved to be minimal, which raised questions for federally-imposed standards in the schools.

In the third section, I discuss the Reading First initiative, which was dogged by political and commercial biases. These setbacks demonstrate how a scientific policy basis does not automatically guarantee scientific implementation. Like the California class-size reduction program, the federal Reading First initiative illustrates the difficulties in applying evidence-based research in education practice. The micropolitical realities that stemmed from the RCT movement resulted in an ineffective federal policy.

In all three cases, the use of the RCT model in education was less successful than in medicine. This was due in part to the nature of the education field, which includes the complexity of educational interventions and their context, and the inability of blinding the trial participants.

I conclude the chapter by discussing the controversy over the Federal Priority of RCT designs in evaluating education projects between 2003 and 2005. The controversy illustrates that the move toward RCTs was not without resistance. In particular, professional organizations such as the American Evaluation Association saw a need to curb this movement because they anticipated negative consequences, such as ideological redirection of funding. Although initially their resistance was unsuccessful, the federal government eventually moved to a more inclusive approach. This example shows how the methodological pendulum, which had swung far from the center, ultimately returned to a less polarized position. I argue that the limited relevance of RCT results for policy makers was the primary motivating factor in steering to a more inclusive position.

1. The Tennessee Student Teacher Achievement Ratio experiment: how a pilot trial influenced state policy and its unintended consequences

Many education researchers regarded the Tennessee Student Teacher Achievement Ratio (STAR) trial in 1985–89 as a prominent example that RCTs would be useful in identifying effective educational interventions. Donald Orlich from Washington State University called STAR “one of the most significant educational research [projects] done

in the U.S. during the past 25 years" (Orlich, 1991, 632). The statistician Frederick Mosteller found that STAR was "one of the most important educational investigations ever carried out and illustrates the kind and magnitude of research needed in the field of education to strengthen schools" (Mosteller, 1995).

STAR findings had a large influence on policy-making as well. Eighteen states including California showed a great interest in the STAR findings and implemented changes based on the findings (Boruch, de Moya, & Brooke, 2002). Even the federal government supported class size reductions; its 1999 federal budget contained 12 billion dollars over seven years for class size reduction purposes (Hoxby, 2000, 1239). The important role that the STAR trial had in research and policy warrants a closer look, raising questions as to what lessons could be learned in the current debate about the value of RCTs.

The context of STAR: Policy and science

The STAR experiment took place in the wake of new federal and state interests in reducing the achievement lag of economically disadvantaged students. For the first time in U.S. history, Congress appropriated federal funds for local school districts serving disadvantaged children through the Elementary and Secondary Education Act (ESEA) of 1965.[11] The goal was to alleviate educational disparities as part of the "Great Society" and "War on Poverty" (Vinovskis, 2002, 122). Until then, education had been funded primarily at the local level through property taxes, levies, and bonds. To stem public criticism that states would squander these newly appropriated funds, ESEA mandated the evaluation of newly implemented education programs. States had to demonstrate that the federally-funded programs promoted educational achievement by submitting reports containing "information relating to the educational achievement of students participating in [federal] programs" (ESEA, sec 206a6). How states developed evidence for educational achievement and what methods of data gathering and analysis they used were left to their discretion.

[11] 89th United States Congress (April 11, 1965). Elementary and Secondary Education Act. Public Law 89-10, 79 Stat. 27.

Around the same time, the psychologists Donald Campbell and Julian Stanley formalized the RCT approach for the social sciences in their monograph, "Experimental and Quasi-experimental Designs for Research" (Campbell & Stanley, 1963; Oakley, 2000). Campbell and Stanley acknowledged the difficulty of applying purely experimental designs to social interventions. They cited the Perry Preschool project in the 1940s as a successful use of RCTs to assess educational outcomes. Campbell and Stanley established a classification system of research designs based on the degree of internal validity, i.e., the ability to attribute causal effect of interventions without bias. The authors argued that any deviation from the RCT would reduce internal validity, and they therefore recommended making every effort to design research studies that adhered as much as possible to the RCT specifications.

In the widely-distributed handbook "Evaluation: A Systematic Approach" (1979), the evaluation theorists Peter Rossi, Howard Freeman, and Sonia Wright made strong use of Campbell and Stanley's approach. They characterized the randomized or "true experiment" as the optimal evaluation design, which reflected prevailing beliefs in the evaluation profession. The authors noted how great their impact was in establishing RCTs as the *sine qua non* in experimental research:

> There is almost universal agreement among evaluation researchers that the randomized controlled experiment is the ideal model for evaluating the effectiveness of public policy. If there is a Bible [sic!] for evaluations, the Scriptures [sic!] have been written by Campbell and Stanley (1966), along with a revised version by Cook and Campbell (1976). The "gospel" of these popular texts is that all research designs can be compared more or less unfavorably to randomized controlled experiments, departures from which are subject to varying combinations of threats to internal and external validity (Rossi, Freeman, & Wright, 1979, 183).

In the Campbell and Stanley tradition, researchers argued that non-randomized trials in education had led to misleading results in the past; they blamed unscientific evaluation designs for compromising policy makers' ability to determine program impact (Brookings Institution, 1966; Rossi, 1972). They concluded that educators had been largely left in the dark as to which classroom and school-wide practices would lead to greater educational achievement.

The "Nation at Risk" report by the National Commission of Excellence in Education called for the need for rigorous education planning (The National Commission on Excellence in Education, 1983). The controversial report argued that mediocrity in education had led the United States into decline in economic competition with other countries, especially the Soviet Union. Based on the report, the U.S. Department of Education released an annual wall chart showing how well each state was doing in education (Vinovskis, 2002, 130). The perceived decline in schooling quality led to state reforms, including more school days, longer days, more tests, stricter graduation requirements, merit pay, and higher teacher qualifications, *inter alia* (Cuban, 1988, 342). The Tennessee education legislation was part of these education reform efforts. However, there was not enough political support for funding class-size reduction across all elementary schools—a feat that would have cost the state $42 million a year (Ritter & Boruch, 1999, 115). The sponsor of the bill, Representative Steve Cobb, suggested a pilot, which reduced the costs from $42 to $5 million a year (ibid., 116). In May 1985, the Tennessee legislature passed a bill authorizing the funding for a "demonstration project … in order to study effects of a reduced pupil-teacher ratio on the achievement of students in public schools."[12] According to the STAR project leader, STAR "was almost a way of putting off spending the money" required for class size reduction (Ritter & Boruch, 1999, 119). This intuition may have been correct because Tennessee never actually enacted the class-size change, despite the positive STAR findings four years later.

Design and findings of the STAR experiment

The version of the bill signed into law made no reference to an RCT (Ritter & Boruch, 1999, 117). The Tennessee Education Agency gave researchers from four Tennessee universities the freedom to design the study according to their ideas. In the spirit of Campbell and Stanley, STAR researchers Finn and Achilles pointed out the confounding issues present in previous correlational studies and non-randomized controlled studies on class size (Finn & Achilles, 1990). Several meta-analyses of class-size reduction had reported only small-scale studies that were primarily observational (Glass & Smith,

[12] An Act to amend the Tennessee Code Annotated (1985), Title 49, Chapter 3, relative to incentives for class size reductions, Section 1.

1979). These analyses triggered an extensive debate over the class-size effect. The STAR researchers saw a need for evidence for substantial and consistent positive outcomes and argued that, on the issue of class size, ultimately "only randomized experiments can provide a definitive answer" (Finn & Achilles, 1990, 558). Thus, they designed STAR as a large-scale, longitudinal RCT involving 11,600 students. Seventy-nine elementary schools participated, which equaled approximately one-sixth of all Tennessee elementary schools.

The STAR authors found that a within-school design was an effective way of controlling for sources of variability between school settings (Finn & Achilles, 1990). Within each school, children entering kindergarten were randomly assigned to three class types: small classes (13–17 students); regular classes (22–25 students); or regular classes with a classroom teacher and a teacher aid. Teachers were also assigned at random to these classes. Every class remained the same type for four years from 1985 to 1989. After the third grade, all students returned to regular-size classes. To alleviate parents' concerns, half of the regular-class students in kindergarten were randomly reassigned to teacher-aide classes when they entered first grade, which confounded the teacher-aid classes with regular classes. The state did not mandate special training for teachers, and schools continued to operate as usual.

The researchers used the Stanford Achievement Tests (SATs) in reading and mathematics as outcome measures. They also administered an academic motivation and self-concept instrument for evaluating small groups of students. A state-developed test for basic skills in reading and math was added beginning in first grade (Finn & Achilles, 1990, 561). Although the data stemmed from individual students, the class was the unit of analysis. The analysis included the calculation of means on each outcome measure for each class, followed with a disaggregation for white and minority students. According to the researchers, small classes in all four grade levels achieved higher test scores, and the effect of class size was similar across school locations (Achilles, report, January 1993, 3). The increased student achievement was both statistically and educationally significant. The "educational" significance derived from the "effect size," i.e., how much the difference was relative to the standard deviation of student achievement. Generally, the

positive effect size for students in small classes ranged between .20 and .27 of the standard deviation (ibid.). Effect sizes favoring small classes ranged from .08 (in K) to .40 (in grade 3) for minority students. The advantage of being in a small class was greater for non-white students than for white students, with average effect sizes of .35 instead of .15. STAR reduced the achievement gap between white and non-white students from 14 percent to 4 percent in the STAR pilot (Finn & Achilles, 1990, 568). The small-class advantage was 1.5 months for reading and 2.5 months for mathematics on the grade equivalent scale (Finn & Achilles, 1990, 567). The study also found that the positive effects from early student participation in small classes remained pervasive two years after the students had returned to regular-size classes, with an effect size of .11 to .22 of a standard deviation (Achilles, report, January 1993, 4–5). The economist Alan Krueger reanalyzed the STAR data and found that a 10 percent reduction in class size for three years raised scores by about 13 percent of a standard deviation (Krueger, 2000).

Discussion of the STAR experiment

The STAR experiment posed several challenges: self-selection of schools, unblinded assignment to treatment and control classes; narrow scope; and no explanation of why reduced class size was more effective. All these challenges most likely affected the generalizability and usefulness of the results. Despite these challenges, undue generalizations about class size reductions were made.

First, the schools did not represent a random sample of Tennessee elementary schools. As in medical trials, STAR used volunteers and exclusion criteria, and suffered from attrition. Only one-fourth of the schools volunteered to participate. Schools with a higher percentage of minority students elected to participate, introducing a self-selection bias. Schools had to serve at least 57 students at a grade level and had to provide sufficient physical space (Achilles, report, January 1993, 2). Thus, small and overcrowded schools were excluded from the STAR experiment (DeAngelis, 2003, 3). Only 55% met the eligibility requirements. Furthermore, approximately 20 percent of students in STAR quit each year, leaving less than half of the original student body by the end of the four years (Hanushek, 1999). Attrition was correlated with disproportionately low performing

students and thus may have positively affected the sample. Self-selection and attrition had implications for the population to which those results could be generalized.

Second, unlike many medical experiments, the assignment to treatment groups was not a blind process. Instead, teachers knew what class size they taught; that is, the actors in the experiment were knowing participants. According to the economist Caroline Hoxby, the participants' knowledge was the biggest flaw of the study and hence tended to work towards achieving desired outcomes. The problem of participant knowledge occurs if participants support one intervention over another, which might influence their way of participation. As Hoxby wrote, "the schools in the class size experiment may realize that if the experiment fails to show that the policy is effective, the policy will never be broadly enacted" (Hoxby, 2000, 1241). Even if schools did not intend to influence policy making, teachers and principals might have expected that students in smaller classes would perform better, and thus their prior expectations may have affected their own participation. Classroom-teacher motivation may have influenced teaching efforts. Also, parents had a clear preference for smaller classes, and some may have lobbied to have their children transferred to the smaller classes (Hanushek, 1999). Pilot projects, such as STAR, may appear to have productivity effects that they might not occur if fully enacted (Hoxby, 2000, 1241). Hoxby claimed that natural experiments were superior because they varied class size, but would not vary incentives. The main advantage would be that participants would not be aware of being evaluated. It would be important that research mimicked the incentives that exist under real policies (Hoxby, 2000, 1281).

Third, the STAR findings were narrow in scope: the study used average class sizes of 16 versus 24. Conclusions whether other degrees of reductions would increase student achievement could not be made. While the overall result suggested a move toward smaller classes, it was unclear what the ideal number would be in terms of pairing the size of the instructional unit and the teaching task (Achilles, report, January 1993, 8). Benjamin Bloom had posed the "two-sigma problem, " which assumed that one-on-one tutoring was most effective for knowledge acquisition and that a decrease in one-student increments also increased the learning effectiveness (Bloom, 1984a; Bloom, 1984b). The question was where the ideal cut-off was for maximizing student learning and minimizing

resource allocation—a question that the STAR experiment could not answer. This problem is similar to the dosage issues of medications in clinical trials. The concept of optimal dosage to maximize benefits and minimize harm translates to education as maximizing student achievement versus minimizing costs. In either case, policy makers must judge how to balance the two sides. As in the Streptomycin trial, one RCT alone could not answer the question about optimal dosage; more trials would be needed.

Fourth, the experiment was not able to answer the question of why and under what instructional conditions small classes work best (Finn & Achilles, 1990, 575). Possible explanations were teacher enthusiasm and satisfaction, more individual teacher attention to students, and higher engagement of students in learning activities. The so-called black box remained in place regarding the processes underlying class-size reduction. Qualitative-interpretive approaches could have helped open the black box.

Despite these challenges, STAR became a "poster child" of RCTs. Hanushek argued that too much emphasis was placed on one single experiment (Hanushek, 1999, 149). One principle of scientific experimentation is that results should be confirmed several times in different contexts before being accepted (Mishes & Rothstein, 2002, 2). This is especially true when the evaluation has not determined the underlying processes of the effects. Even if the STAR results had been valid, policy conclusions were drawn that go beyond what STAR alone could show. In Hoxby's mind, STAR interpreters extrapolated the results unduly (Hoxby, 2000, 1241).

The STAR results, however, did not result in the intended policy change, such as state-wide class-size reductions. Tennessee lawmakers implemented smaller classes for K–3 students in merely 17 small rural districts and, later in 1992, mandated class sizes of a maximum of 20 for K–3 grades—not the 16 student maximum used by STAR (Ritter & Boruch, 1999, 119). According to the 2002 book "The Class Size Debate," the STAR experiment did not ultimately settle the debate over class size (Mishes & Rothstein, 2002). As author Jennifer King Rice noted, even with the STAR experiment, the literature had offered little closure or clear directions for policy makers who considered investments in smaller classes (King Rice, 2002, 90). Although the quality of the

statistical results of STAR was high, divergent interpretation of the results and of their implications arose. The debate illustrated how the same evidence could lead to different conclusions. On the one hand, researchers confirmed the intuition of teachers that class size mattered, and some argued that the added cost was worthwhile. On the other hand, some researchers argued that the gains would not justify the costs. John Folgers, one of the key STAR investigators, argued that the 25 percent added cost in educating elementary students would not justify the achievement gains. He referred to other school reform efforts, such as "Success for All," that were more cost-effective than reducing class size. Even the original sponsor of the bill, Steve Cobb, found the results disappointing. Given the resource constraints, Cobb asked: "How much for how much?" (Ritter & Boruch, 1999, 120) Eric Hanushek lamented that the costs of broad class-size reductions were rarely compared to other potential uses of funds (Hanushek, 1999, 144), which would have followed the tradition of the medical comparative effectiveness research (cf., chapter two). Hanushek argued that grades 1 to 3 did not benefit much from class-size reduction, and thus STAR as implemented was an inefficient use of funding (Hanushek, 2002). Hanushek concluded that class size reduction was a political decision based on voter support and preconceived policy proposals, but that it was not based on scientific evidence (Hanushek, 2002, 62). Hanushek's method of meta-reviewing prior research was controversial, however, and his criticism should only be taken *cum salo granis*.

These debates lead to questions about how much evidence policy decision makers need before they apply findings to help improve schooling (Zigler, 1992). There is indication that class size reductions help, but two questions remain surrounding the implementation of smaller classes: With what costs, and under what circumstances is class-size reduction advised? The STAR experiment could not readily answer these questions. More interpretive-observational and cost-effectiveness approaches were needed to adequately address these questions.

California's Class Size Reduction initiative and its unintended consequences

STAR-results inspired California lawmakers who were concerned about their students' performance on the 1994 National Assessment of Educational Progress (NAEP). Fourth

graders ranked last place in reading among 39 states that participated in the NAEP (Bohrnstedt & Stecher, 2002, 4). At that time, California elementary schools had the largest class size in the country, averaging 29 students. A task force assembled by the California Department of Education called for smaller classes—a move strongly favored by teachers' unions and parents alike. New tax revenue from the dot-com boom allowed the implementation of these changes. The state legislature passed Senate Bill 1777 in 1996, which aimed at cutting class size from an average of 29 students to a maximum of 20 (Mishes & Rothstein, 2002, 3). School districts that chose to participate in the program, which was voluntary, received about $850 for each K–3 student enrolled in a small class. In addition to funding at the student level, the state also offered facility grants (Stecher & Bohrnstedt, 2002, 3). The school districts implemented the Class Size Reduction (CSR) program quickly, with urban, high-minority districts lagging behind due to facility constraints (Stecher & Bohrnstedt, 2002, 5). The state of California phased in the initiative over three years to K–3 classes, starting with first grade.

The California Education Department established an evaluation consortium, which included the American Institutes of Research and the RAND Corporation. They attempted to retroactively determine whether CSR had increased student achievement. Their challenge was to analyze observational data. Although STAR was an RCT, the statewide implementation of CSR did not allow for a randomized control study, and comparisons could only be made to previous years. During these years, California also implemented other policy changes, such as a stronger system of accountability. Therefore achievement changes could not be attributed solely to class size reduction. California used the Stanford Achievement Test 9, initiated a few years prior. The evaluators found an increase in achievement across the board. However, the evaluators determined that the relationship between average achievement scores in the new statewide testing program and the CSR initiative was inconclusive. Researchers expected test score increases after the implementation of a new test (Stecher & Bohrnstedt, 2002).

Despite the missing evidence on academic improvement, the evaluators reported that the CSR initiative in California was an enormously popular program among elementary parents and teachers (Stecher & Bohrnstedt, 2002, 8). Local educators and parents

seemed to value reduced class sizes for reasons beyond improved achievement as measured by statewide test scores (Bohrnstedt & Stecher, 2002, 8).

The evaluation reported several shortcomings and recommendations. The evaluators were not able to determine what type of changes in classroom teaching would be needed to maximize the benefits of the reform. Therefore, they suggested more research determining which classroom practices were most effective in small classrooms, and whether these practices differed from best practices in larger classes (Stecher & Bohrnstedt, 2002, 11). The evaluators recommended calculating the real cost of CSR and that the state provide adequate resources for CSR (Bohrnstedt & Stecher, 2002, 9). They also recommended improving the effectiveness of the CSR program by integrating and aligning it with other reforms, such as the state's new standards-based policies (Stecher & Bohrnstedt, 2002, 9). Based on the evidence from the Tennessee STAR experiment, they recommended targeting additional resources to urban, high-minority schools (Stecher & Bohrnstedt, 2002, 10). They suggested creating and evaluating pilots with smaller class sizes (e.g., 15 students). Moreover, the state could compare the effectiveness of the current CSR program with alternative CSR designs by allowing a small number of school districts to use their CSR funding to create randomized trials of other small class size arrangements (Stecher & Bohrnstedt, 2002, 11). The CSR evaluators promoted the idea of comparative effectiveness research in class-size evaluation.

Discussion of education policies based on pilot experiments

The Tennessee STAR project directly inspired the California Class Size Reduction initiative (Stecher & Bohrnstedt, 2002, 3). However, there were significant differences between STAR and California's CSR, which may explain the differences in impact. The CSR evaluators advised against regarding CSR as a replication of the STAR experiment; they advised judging CSR on its own terms (Stecher & Bohrnstedt, 2002, 3). A key difference was that the California initiative was not a pilot project, but a statewide program. Whereas STAR involved 6,500 students per year, CSR served 1.8 million students statewide. As a result of CSR, the number for K–3 teachers in California increased by 46 percent. The demand for credentialed teachers outpaced their supply and

led to an inequitable distribution of credentialed teachers (Stecher & Bohrnstedt, 2002, 6). The number of teachers with less than three years of experience increased significantly from one out of five to one out of three teachers (Stecher & Bohrnstedt, 2002, 40). The proportion of teachers without full credentials increased from 1.8 percent before the initiative to 15 percent in 2000–01 (Stecher & Bohrnstedt, 2002, 2). In schools with economically disadvantaged students, on average 21 percent of teachers were not fully credentialed in 2000–01. In contrast, all of the new teachers hired for the STAR experiment in Tennessee were fully certified (Finn & Achilles, 1990). Another possible difference may have stemmed from teach education and training, given that many teachers had graduated from universities within their respective state.

Whereas STAR reduced class sizes from 24 students down to an average of 15, California reduced class size of 29 to 20 students. The STAR research could not determine the gradation level necessary to improve academic achievement. It only provided answers to a decrease of 7 students from a class size of 24. It was unclear whether a decrease of 9 students to 20 students had a similar effect, as in the case of California. STAR had confirmed through prior research that a class size of 15 students was more effective than a class size exceeding 20 students (Glass & Smith, 1979). The student populations between STAR and CSR students also differed significantly: California served a much more ethnically and linguistically diverse student population than did Tennessee.

Another important difference was the funding levels between STAR and CSR. Whereas Tennessee carried the full cost of teachers, California only gave an $850 supplement to each individual student. This amount did not cover the cost of the added hires. Schools and districts had to reallocate funds from other programs to support the implementation for CSR (Stecher & Bohrnstedt, 2002, 2). Two-thirds of districts reported that the state reimbursement for CSR was insufficient to cover actual district costs. Funding was reallocated from music and arts, gymnastics and sports, professional development, computer labs, libraries and after-school programs (Stecher & Bohrnstedt, 2002, 8). In addition, many California schools did not have the necessary space available for the CSR program. They had to reallocate space such as libraries, stages, and other rooms used for

various purposes. CSR was a more realistic scenario than STAR, because some schools had to reallocate existing funds rather than benefit from new added funds. Had CSR allowed for randomization across schools, this condition would have allowed for comparative effectiveness research. For example, one could have answered the question as to whether financial resources would have been more effectively used in reducing class size rather than, say, providing professional development for teachers.

The STAR experiment compared class-size reduction to regular classrooms. It did not concurrently compare multiple interventions which would require the same amount of funding, such as providing a fixed amount for class size reduction versus for increasing the school year. Therefore, comparative cost-effectiveness analysis was not possible, which would help policy makers answer the question how to spend a fixed amount most effectively—in light of limited and competing public budgets.

Furthermore, calculations of the benefits of class-size reduction, given the estimated changes in future earnings, were difficult. For instance, given that all Tennessee graduates would be better educated if the pilot were to be implemented statewide, the earning levels might not change much. Expanding the program to a larger population may not result in the expected earnings due to other variable changes, such as increased supply of better-educated workers.

The California initiative illustrated how statewide programs differ from pilot projects. Had Tennessee implemented their pilot experiment statewide, a shortage of teachers and space would have been anticipated. The experiment excluded overcrowded schools. The state did not provide additional funding for facilities in their pilot experiment. Therefore, it is unclear how STAR would have fared as a statewide initiative. California was inspired by Tennessee, but did not implement STAR findings. One reason for this was that California's average class size was five students higher than Tennessee's. A reduction to 15 students would not have been financially feasible, as costs of education would have increased by maybe as much as another billion dollars and would have created an even larger shortage of teachers.

The examples of Tennessee and California illustrate the difficulties with extrapolating findings and applying them from one setting to another. Just the fact that California serves a much larger and more diverse student population than Tennessee makes a comparison difficult. The philosopher Nancy Cartwright pointed out that the California initiative produced unintended negative consequences, concluding that generalizations from experimental settings, such as STAR, might end up in deficient policy solutions (Cartwright & Munro, 2010). Cartwright highlighted the difficulties of generalizing from RCTs in the way STAR had: Too often, the temporal-spatial and contextual constraints of an RCT barred generalizations to different geographical and temporal settings.

Unfortunately, California decided not to implement a pilot study. However, a pilot study would not guarantee that findings would produce meaningful results. For one, a pilot study would not have triggered resource problems such as teacher shortage, as STAR illustrated. Therefore, extrapolating findings from a pilot intervention to a large-scale program is not straight forward. The problem of generalizability from a pilot project to the general population had been illustrated in the 1950s by the Salk trials (cf. chapter 2). The oversight of the vaccine production process was less stringent, given the exponential increase in demand for that vaccine. As a result, the vaccine products used in the Salk trial and in the general distribution differed. This problem of re-applying pilot findings is even larger in educational systems with more factors to consider. The composition of a drug is much less complex and subjective than the composition of a teaching environment. Drugs can be produced more quickly than teachers, school buildings, or other infrastructure. A vaccine can be administered in minutes, whereas a teaching intervention takes months. Facility requirements are different as well. All these additional factors in education make the application from a select pilot population to a general population much more difficult in education than in medicine.

Note that within the field of education, class-size reduction is a comparatively simple intervention, because only the student-teacher ratio is changed. No immediate changes in the curriculum, teaching methods, or teacher training took place in the STAR experiment. The following discussion of reading interventions demonstrates the greater challenges of

generating generalizable evidence for more complex programs, such as reading interventions.

2. The National Reading Panel's RCT-guided standards of evidence: how the medical model influenced research standards in education

One of the primary debates in education policy during the Clinton and George W. Bush administration was how students should be taught to read. As with the debate over class-size reduction, researchers and policy makers looked to RCT findings to settle the debate. In contrast to a class-size intervention, a reading intervention is more complex, because it requires a substantial change in the student-teacher interaction: Different instructional approaches, new teaching materials, and professional development for teachers may be necessary ingredients to implement a reading program. There are also many different ways for constructing a curriculum, teaching reading, or training teachers. It proved to be challenging to determine what type of reading programs would improve reading achievement most effectively.

Reading First was a product of the increasing culture of accountability in education as enforced by statewide student achievement tests. In 1994, the newly implemented National Assessment for Educational Progress (NAEP) found that 69 percent of African American and 64 percent of Hispanic-American fourth graders tested below basic reading levels set by NAEP.[13] The statistics sent a message to policy makers that a reading problem existed among America's school children (Song, Coggshall, & Miskel, 2004, 446). The concern for the problem of reading ability spanned partisan divides (Song et al., 2004, 449). In his 1997 State of the Union address, President William J. Clinton suggested a "national crusade for education standards" (Clinton, February 4, 1997). He stated that forty percent of eight-year-old students could not read and, therefore, that "we must do more to help all our children read." To respond to the so-called "national reading crisis," in 1998 the U.S. Congress passed the Reading Excellence Act (REA).[14] As part

[13] National Report Card, retrieved from www.http://nces.ed.gov/nationsreportcard.
[14] 105th United States Congress (October 21, 1998) Reading Excellence Act. A part of the Omnibus Consolidated Appropriations and Emergency Supplemental Act. Public Law 105-277. 112 Stat. 2681.

of the Omnibus Consolidated and Emergency Supplemental Appropriations Act, the U.S. Congress allocated $260 million for competitive grants to state education agencies for local low-income schools to implement scientifically based reading instruction (Sweet, 2004, 21). REA included a definition of scientifically based research (SBR), synthesized from various sources by staff from the House Education and Workforce Committee (Eisenhart & Towne, 2003). The final bill defined SBR as "systematic, empirical methods that draw on observation or experiment" (112 Stat. 2681-393). The law regarded experimental *and* observational methods as scientific. Conversely, the federally mandated National Reading Panel's report of 2000 limited the concept of "scientific" to experimental and quasi-experimental research, which then influenced the 2001 NCLB law.

The National Reading Panel's utilization of the RCT-guided medical model

In 1997, the U.S. Congress mandated the establishment of a National Reading Panel (NRP) "to assess the status of research-based knowledge, including the effectiveness of various approaches to teaching children to read" (NICHD, 2000, 1-1). Congress asked the National Institute for Child Health and Human Development (NICHD), part of the National Institutes of Health (NIH), to convene such a panel.[15] NICHD had produced reviews on reading difficulties in the past (Coles, 2003, 12), and NIH had a long history of using scientific panels to resolve empirical controversies. It was the first time in history that such a panel was created for education research (Shanahan, 1999). NICHD was a natural choice in convening and managing such a panel for an education controversy. The charge left open what to base the assessment on.

When the NRP bill was introduced in the Senate, Duane Alexander, NICHD director and medical doctor by training, testified with the following:

> "I think that it is important to point out that our [NICHD's] intensive research efforts in reading development and disorders is motivated to a great extent by our seeing difficulties learning to read as not only an educational problem, but also a major public health issue. Simply put, if a youngster does not learn to read, he or she will simply not likely to [sic] make it in life" (Alexander, testimony, June 19, 1997).

[15] The justification for increasing funding for NICHD in Congressional Record House (November 7, 2007). Conference report on H.R. 2264, Department of Labor, Health and Human Services, and Education, and related agencies appropriations H10230.

Alexander made the case that a reading problem is a public health issue. In a legislative hearing, Reid Lyon from NICHD argued, that "NICHD considers reading failure a national public health problem," (Lyon, testimony, October 26, 1999). In another instance, he argued that "NICHD considers that teaching and learning in today's schools reflect not only significant educational concerns but public health concerns as well" (Lyon, testimony, March 8, 2001). Both Alexander and Lyon from NICHD compared reading problems to health problems, and reading interventions to health interventions.

Similarly, when NRP chair Langenberg presented the NRP report to Congress, he testified that:

> "No physician would normally subject a patient to a treatment or a drug whose efficacy had not been proven in rigorous scientific testing, and we should expect no less of a teacher subjecting a student to the curricular content or a teaching methodology" (Langenberg, testimony, April 13, 2000).

Langenberg drew parallels between education treatments and medical treatments, teachers and physicians, and education testing and medical testing. From this perspective, it made sense to use the methodological standards of medicine also in the field of education.

In NRP's inaugural meeting in April 1998, one agenda point was the "review of models of methodological approaches" for analyzing research (NRP meeting minutes, April 24, 1998). Alexander described two medical models for reviewing and evaluating research findings—the Cochrane Collaboration Model and the Best Evidence Synthesis Model. He emphasized their possible relevance for analyzing literature outside the medical field. Both models put the RCT at the top of the evidence hierarchy, in line with the approach of Evidence-Based Medicine (cf., chapter 2).

The NRP explicitly adopted the medical model to evaluate research. Like medical research, the panel developed a set of rigorous methodological standards for searching, selecting, and analyzing research. In the panel report's own terms: "The evidence-based methodological standards adopted by the Panel are essentially those normally used in research studies of the efficacy of interventions in psychological and medical research"

(NICHD, 2000, 5). In a similar vein, the panel's chair, Donald N. Langenberg from the University of Maryland, stated the following during the official presentation of the NRP report to the Committee on Appropriations on April 13, 2000:

> "I think the most important thing the Panel did was what it did next, and that was to develop a set of rigorous methodological standards to help them screen the research literature relevant to each topic. Those standards are essentially those normally used in medical and behavioral research to assess the efficacy of medications, medical procedures, or behavioral interventions" (Langenberg, testimony, April 13, 2000).

As evaluators had determined in the case of the medical model, the Panel determined the RCT to be the best evaluation design in education. Following the RCT, quasi-experimental methods were judged to be a merely "acceptable" standard to answer causal questions. In contrast, the panel found descriptive and correlational research ill-suited for making any causal claims (NICHD, 29). Qualitative studies would primarily deepen the understanding of *how* things worked (Shanahan, 2004, 244). The report's addendum stated accordingly:

> "The highest standard of evidence for such a [causal] claim is the experimental study, in which it is shown that treatment can make such changes and affect such outcomes. Sometimes when it is not feasible to do a random experiment, a quasi-experimental study is conducted" (NICHD, 2000, 29).

Therefore, the panel included in their review process only reading programs that used an RCT or a quasi-experimental design with a control group (ibid., 5); they automatically screened out evaluation designs using interpretive and qualitative methods. The Panel thus distinguished three grades of evidence:

TABLE 5: Hierarchy of evidence of the National Reading Panel (2000)

Grade	Type of Evidence
Highest standard of evidence	Experimental studies [=RCTs]
Acceptable standard of evidence	Quasi-experimental studies
No claim of evidence	Correlational and descriptive studies

Source: Adapted from the NPR summary report (2000, 29)

In their attempt to review existing literature on effective reading instruction, the reading panel's underlying assumption was that a dichotomy existed between experimental studies, on the one hand, and observational and qualitative studies, on the other hand. The

panel member Timothy Shanahan granted that medical researchers had also used correlational evidence for some determination of causal claim (e.g., the relationship between cigarette smoking and lung cancer). These correlations would not serve, however, as the only evidence for lung cancer. Animal studies, for example, had experimentally tested the causal connection between smoking and lung cancer as well. Shanahan also argued that those medical correlations were produced in much more "sophisticated" ways than traditionally done in reading research (Shanahan, 2004, 247). He justified the exclusion of qualitative studies on the grounds that, without persistent observations, these studies were generally not rigorous, and only exploratory in nature (Shanahan, 2004, 243).

Alexander argued that the methodology developed for the literature analysis was a major contribution to the field of literacy research (NRP meeting minutes, October 19, 1998). The NRP saw their emphasis on how to systematically construct evidence as a major advancement over previous reviews such as the National Research Council's (NRC) review on summarizing effective reading research (Snow, Burns, & Griffin, 1998). The NRC consensus report "Preventing Reading Difficulties in Young Children" did not include clear guidance on selecting methods for their review. According to the NRP member Timothy Shanahan, reading policy and practitioners' choices had been an "idiosyncratic affair—with each researcher or practitioner making his or her own decisions about the implications of research" (Shanahan, 2004, 235).

In 2000, NRP released their review "Teaching Children to Read—An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction" (NRP, 2000). They concluded that instruction in phonemic awareness, phonics, fluency, vocabulary, and comprehension—as well as increased teacher education—all improved a student's reading achievement. The panel found insufficient evidence to conclude positive effects for reading technologies and for strategies of encouraging children to read, and thus they recommended conducting more research on these topics until they could make a final conclusion.

Discussion of the National Reading Panel's findings

Although the National Reading Panel attempted to use the medical model for producing scientific standards and findings, they struggled with adapting it to the field of reading research. First, they selected the review topics without clear guidance; second, they screened out 99 percent of the existing reading literature; third, they struggled to make meaningful recommendations based on the few studies; and finally, they were not protected from commercial bias.

Difficulties in applying the medical model to the education model stemmed from inherent differences between the two fields themselves. Medical interventions are often discrete. The Tuberculosis trial of 1948, for example, only involved Streptomycin as an active drug and addressed a biological problem, i.e., the spread of a bacterium (cf., chapter 2). In education, however, an intervention comprises many components that are often difficult to standardize. The teacher has a special role, not only in deciding on the particular treatment like a physician, but in directly influencing a child's learning. There are infinite amounts of variations in teaching how to read.

First, NRP selected the review topics without clear guidance. Granted, their task was almost unmanageable. Whereas Drug Efficacy Study in 1969 had struggled in reviewing 300 distinct chemical formulae (National Research Council, 1969; cf., chapter 2), the National Reading Panel was faced with more than 100,000 studies due to the heterogeneity of reading interventions. To make the task manageable, the panel had to limit their work to a few topics. This process of selecting topics was less systematic and much more qualitative-interpretive in nature than the establishment of the evidence standards. The meeting archives demonstrate that the panel considered approximately 30 topics for review at the beginning (Shanahan, 2004, 237). The panel then voted which topics to examine and settled on the eight topics that received the highest number of votes, but "those selections were not uniformly agreed to" (NRP meeting minutes, October 29, 1998). These topics resembled the areas identified by the 1998 National Research Council report, which the NRP attempted to supersede (Snow, Burns, & Griffin, 1998). The panel's focus on phonemic awareness as an important component, for example, was a claim based on the convictions of just a few panel members (NRP

meeting minutes, September 10, 1998). Later, the panel identified an additional twelve areas to possibly include in the review, but then dropped their proposal due to limited time (NPR meeting minutes, November 9–10, 1998). What subgroups and topics were selected in the first place, however, influenced the final panel findings. A more systematic selection process would have made the final pillars of literacy instruction less accidental in their choice.

The panel member Joanne Yatvin pointed out that many areas were left out, such as research on the relationship between reading and writing, language development or the understanding of print (Yatvin, 2000). Other researchers also had concerns about excluded topics, especially the influence of writing, motivation, and home experiences on reading (Shanahan, 2004, 240). Granted, the panel report stated that it did not consider the chosen topics to be the only topics of importance in learning to read; the omission did not mean that those topics were irrelevant for reading instruction (NICHD, 2000, 3). The panel simply lacked the resources and time to study all possible issues in reading (Shanahan, 2004, 241). This statement is of high concern given that the choice of topics would influence the nature of the findings. Had the panel picked other topics, the report would have most likely found pillars of reading instruction different from the ones recommended. Furthermore, some studies were not meta-analyzed because they were so diverse that the panel "lacked the time to provide the kind of detailed analysis that they deserved" (Shanahan, 2004, 253). Resource and time constraints are common in a policy context. This situation raises the question, however, of whether additional time and resources would have changed or would have just refined the final findings of the review. If time and resource constraints affected the findings, the panel could only produce recommendations, constrained by and subjected to contextual demands. Such constraints might lead to less than "objective" findings.

Second, by using the medical RCT model, the panel's review was reduced to a small percentage of existing studies. Their focus on (quasi-)experimental studies pruned the number from 100,000 possible studies to approximately 320 studies, which amounted to less than one percent of the eligible studies. The fact that most qualitative studies seemed flawed did not justify their untested exclusion. In the end, the process excluded a large

branch of educational research prematurely. Although the (quasi-) experimental lens made the reading panel's task manageable, the large amount of excluded literature might raise questions about whether a different selection process may have changed the panel's findings. One problem was that RCTs were unevenly spread across the range of reading interventions. Whole-language and balanced reading approaches had undergone fewer RCTs than phonics programs (Sweet, 2004, 29), a problem pointed out by *The Economist* (Economist, February 28, 2002, 73). These RCT-sparse programs were automatically excluded from the review process. The panel did not recommend interventions with insufficient RCT evidence, which did not mean that these interventions were ineffective; they were simply unproven. One important question for policy makers is: What evidence and how much of it do they need for promoting a policy intervention? Does it have to be experimental evidence of the most rigorous nature? Or would it be sufficient to relax the definition of evidence?

Third, the NRP was unable to identify why certain reading strategies were successful. Such information was necessary for making evaluation results relevant for wider use. They reviewed RCT studies that primarily compared average effects of final outcomes, but not why those took place. Studies under review sometimes varied several variables at once. In one case, the treatment group had a lower student-teacher ratio than the control group, so the study might have demonstrated effectiveness in reducing class size rather than in a specific reading intervention (Coles, 2003, 47). In another study, the treatment group received individual tutoring on learning word skills and writing, while the controls were tutored as a group without learning word skills and without writing (Coles, 2003, 81). The different intervention outcomes could be attributed to one of several factors, and the reviewed evaluation study could not identify which causal claims were correct. In these cases, the so-called black box of the underlying processes of reading comprehension was not opened. The inclusion of more qualitative studies might have answered questions about why certain types of instruction were more successful than others. As pointed out in chapter 2, this problem of black-box evaluations also accompanies medicine. RCTs drug studies cannot determine why a drug is effective. However, since the interventions are much more discrete, i.e., a single drug is tested, the question about why a drug works is less relevant. Since reading interventions usually

combine many components, it would be important to identify the underlying process of how they work in order to successfully replicate them.

A related criticism about relevance was that the panel ignored contextual factors, which made their findings less relevant. The panel member Yatvin complained that in most cases they did not consider "school and classroom realities" that would make certain types of instruction difficult to implement (Yatvin, 2002). She argued that many research findings were abstract, and it was unclear whether they were immediately useful in the classroom. Austin Bradford Hill had already recognized the problem that RCT findings might not be easily applied to the "general run of patients" (cf. chapter 2).

Despite their insistence on using an experimentally rigorous approach, the NRP made use of qualitative-interpretive reasoning in their review process. Meta-analysis itself required qualitative reasoning, especially when the materials were heterogeneous, hard to compare, and scarce: "Where there were too few studies that satisfied the panel's criteria to permit a meta-analysis, the panel made a decision to conduct a more subjective-qualitative analysis to provide the best possible information about an instructional topic" (NICHD, 2000, 5). Contrary to critic Elaine M. Garan's claim that the studies themselves were of qualitative design (Garan, 2002), it was the meta-analysis itself that was of a qualitative-interpretive nature. The panel relied on interpretive skills to come to its conclusions. This was especially the case in the area of reading comprehension, where the reviewed study findings were not homogeneous (NICHD, 2000, 5). The NRP did not explain, however, how they utilized qualitative reasoning to arrive at their conclusions.

Finally, some argued that the findings were predetermined—a criticism similarly levied at the commercial bias in medical evaluations (cf., chapter 2). The panel member Joanne Yatvin found that "all the scientist members held the same general view of the reading process" and that they agreed on a "hierarchy-of-skills model" of learning to read without debate (Shanahan, 2004). Another widely-held reading model was the constructivist or holistic view of reading, which only one of the fifteen panel members supported. NICHD's preference for "experimental" researchers may have stemmed from the medical context, within which NICHD operated (such as being an entity of the National Institutes

of Health). Yatvin, Garan, and others pointed out some panel members' ties to commercial reading programs such as McGraw-Hill's Direct Instruction and thus had vested interests in the outcome of the report, professionally and financially (Garan, 2002, 77). Garan drew the lines of influence among panel members, the Bush Administration, and the publisher McGraw-Hill, who had been a major publisher of reading textbooks. Harold McGraw, the chairman of McGraw-Hill, had close ties to Bush's Texas 1994–2000 governorship and was on the board of the Barbara Bush Foundation, and McGraw-Hill authors helped guide the Texas reading initiative.[16] McGraw-Hill was able to increase its market share to 37 percent of the Texas market in K–3 literacy textbooks and enter other states such as California, where his textbooks served half of the schools under the Reading First program (Garan, 2002, 80; Coles, 2003, 77). Although these financial-political ties would not automatically lead to biased evaluations, they coincided with the panel's findings.

The NRP's influence on the What Works Clearinghouse

The National Reading Panel's summary report became a model for research syntheses on other topics (Olson & Viadero, Education Week, January 30, 2002). In 2002, the U.S. Department of Education established the What Works Clearinghouse (WWC), a public-private partnership institution to be a source and tool "to provide educators, policy makers, researchers, and the public with a central, independent, and trusted source of scientific evidence about, what works in education" (Institute of Education Sciences, May 17, 2003). WWC attempted to promote the "coordination, development, and dissemination of scientifically valid research in education" (Cook & DeMets, 2008, 8). The WWC was designed as a tool similar to the review process of the Federal Drug Administration. In the legislative hearing on the authorization of DOE's research arm, Grover (Russ) Whitehurst, the Assistant Secretary for Educational Research and Improvement, envisioned the revolution that had taken place in medicine in the last fifty year to "take substantially less than 50 years to get it accomplished in education" (Whitehurst, testimony, June 25, 2002). The proposed WWC would help in this effort.

[16] While Texas students increased their passing rate on the statewide reading test, they did not on the National Assessment for Educational Progress in reading (Coles, 2003, 117). This could be an indication that Texas was not the poster "education miracle," but that schools might have taught to the statewide multiple-choice test to improve reading scores.

According to the frequently asked questions and answers on the early WWC website, NCLB moved the testing of educational practices toward the medical model (U.S. Department of Education, August 2, 2002). The WWC set clear criteria for including and excluding primary evaluations in the review. Similarly to the NRP's review criteria (cf., TABLE 5), the handbook distinguished three grades of evidence, summarized in the following table:

TABLE 6: Hierarchy of evidence of the What Works Clearinghouse (2008)

Grade	Type of Evidence
Meets evidence standards (strong evidence)	Well-designed and well-implemented RCTs with low attrition
Meets evidence standards with reservations (weaker evidence)	RCTs with high attrition[17] or designs with equivalency and low attrition
Does not meet evidence standards (insufficient evidence)	Designs with equivalency and high attrition or designs without equivalency (regression discontinuity and single case studies not determined yet)

Source: Adapted from the What Works Clearing House Handbook (2008)

The handbook identified RCTs as the only type of evidence to meet the highest level of evidence: "Currently, only well-designed and well-implemented randomized controlled trials (RCTs) are considered strong evidence, while quasi-experimental designs (QEDs) with equating may only meet standards with reservations." [18] The core concern was the equivalence of a comparison group. The evidence hierarchy automatically excluded non-experimental evaluations, i.e., evaluations without equivalent comparison group.

The DOE's WWC has functioned as a review filter for (quasi-)experimental evaluation in education since 2004. In general, the WWC has been a catalyst in distributing RCT findings in theory, rather than acting as an authority on education policy at the state and local level. The WWC may not have influenced educational practices as strongly as they had originally hoped to. The U.S. Government Accountability Office (GAO) surveyed

[17] High attrition is defined as an effect size of 0.05 of a standard deviation or more on the outcome variable.
[18] Quasi-experimental designs make the assumption that the intervention and comparison groups are equivalent, based on observable characteristics. Their equivalency may not hold due to unobserved characteristics. Therefore, quasi-experimental designs provide evidence with reservations (i.e., weaker evidence).

several states, their school districts, and schools on their utilization of the WWC reviews (U.S. Government Accountability Office, July 2010). One core problem was that WWC only found a few programs to be effective when employing the experimental evidence hierarchy. Therefore, school districts and education practitioners did not find the reviews useful for their curriculum planning. The experimental legacies of No Child Left Behind, the USDOE experimental priority, and the WWC reviews have made marks in the evaluation discourse in the United States and may also have fanned RCT movements in European countries such as the United Kingdom.

The NRP members could not have anticipated the report's significance in the creation and implementation of national policies. In fact, just when the NRP had started its work in 1999, panel member Timothy Shanahan stated:

> "Some of my colleagues think that the National Reading Panel is a kind of slippery slope—they fear that it creates a dangerous precedent that will be hard to live with. The NRP's reliance on quantitative, experimental results, they sometimes argue, makes it possible to conclude that only certain types of research evidence have value and, therefore, the federal government might decide to fund only work based on such research paradigms. I think this argument is far-fetched. In any event, the government has never explicitly set such methodological limits in the past, even in fields such as medicine, where research findings have long been used to establish standards of practice" (Shanahan, 1999).

Shanahan's cautionary statement illustrates that the NRP report's influence was quite unpredictable. Although the panel did not set methodological limits for federally-funded education research, policy makers used its work as an exemplary way to generate evidence. In this way they justified the new demands for experimental research design.

3. Reading First: how the call for evidence-based research materialized in practice

The Reading First initiative was a central piece of the No Child Left Behind (NCLB) legislation.[19] Reading First attempted to utilize the evidence provided by the National Reading Panel and implement their findings in the classrooms. However, the ideal of scientifically based reading instruction manifested itself as commercial bias.

[19] 107th United States Congress (January 8, 2002). No Child Left Behind Act of 2001; Public Law 107-110, 115 Stat. 1425.

Scientifically based research and government control: the context of the Reading First initiative

NCLB, the reauthorization of the Elementary and Secondary Education Act of 1965,[20] focused on, among other things, the achievement lag experienced by economically disadvantaged and minority students. The overarching goal of Reading First was to improve students' reading achievement. Like many preceding initiatives, the program intended to close the achievement gap in reading among schools serving economically disadvantaged students. The stated goal was to help these low-income schools adopt strategies "that have been proven to prevent or remediate reading failure," grounded on scientifically based reading instruction. As mandated by the NCLB (Title I, Part B, Subpart 1), Reading First provided grants to states for establishing reading programs in kindergarten through third grade. The legislators expected that these reading programs would help students make significant progress on standardized tests toward state-defined reading proficiency by third grade.

The implementation of Reading First cost five billion dollars in its five years of operation from 2002 to 2006. To justify these expenses, states had to demonstrate that their programs were anchored in scientifically based research. NCLB and its Reading First program regarded scientific evaluations as a means to uncover ineffective education programs, contributing to the achievement gap. Through its legislation, the federal government attempted to fund only scientifically based education programs. What NCLB meant by scientifically based research (SBR) was initially unclear. At one place, the law defined SBR as

> "research that is evaluated using experimental or quasi-experimental designs in which individuals, entities, programs, or activities are assigned to different conditions and with appropriate controls to evaluate the effects of the condition of interest, with a preference for random-assignment experiments, or other designs to the extent that those designs contain within-condition or across-condition controls." (Sec. 9101)

Within NCLB, the Reading First section did not include this experimental definition of what counts as SBR in education, but instead refers to "empirical methods that draw on

observation or experiment" (Sec 1208). Based on the different definitions in the NCLB legislation, the U.S. Department of Education (DOE) commissioned the National Research Council to publish a report about the meaning of the term "scientifically based research." The report concluded that the experimental nature of science in education and of science in other fields was substantially similar:

> "Ultimately, we failed to convince ourselves that, at a fundamental level beyond the differences in specialized techniques across the individual sciences, a meaningful distinction could be made among social, physical, and life science research, and scientific research in education. At times, we thought we had an example that would demonstrate the distinction, only to find our hypothesis refuted by evidence that the distinction was not real" (Shavelson, Towne, & Committee On Scientific Principles for Education Research, 2002, 51).

The USDOE used the more rigorous interpretation of scientifically based education in their Reading First implementation, based on NRP's medical model. According to a USDOE guidance document, reading instruction was "an area where some of the best and most rigorous scientifically based research is available" on "what works" (U.S. Department of Education, 2002, 1). USDOE further stated that "program effectiveness has been shown through an experimental design that includes experimental and control groups created through random assignment or carefully matched comparison groups" (U.S. Department of Education, 2002, 44). This call for experimental methods influenced the way states implemented Reading First.

Reading First was a tricky initiative. The 1979 Department of Education Organization Act upheld the state and local sovereignty principle that did not allow the federal government to interfere with state-level and local decision-making. [21] This meant that USDOE officials were forbidden to exercise any control over the curriculum (DEOA, 3403b). However, the federal government put the Reading First initiative in place to improve reading achievement via reading initiatives. How could the federal level not exercise any control, while at the same time counseling on evidence-based reading programs? Michael Sweet stated that it was unclear how to promote findings of scientific research without imposing the use of a specific textbook (Sweet, 2004, 25). Reading First

[20] 89th United States Congress (April 11, 1965). Elementary and Secondary Education Act. Public Law 89-10, 79 Stat. 27.

and No Child Left Behind were considered the "most prescriptive of any federal education law to date" and "pretty close to the edge of what the law allows" (Education Week, September 7, 2005). The U.S. Office of Inspector General confirmed the suspicion that the federal government overstepped their role in implementing reading programs in classrooms when upholding the state and local sovereignty principle.

The Inspector General's criticism of the implementation of evidence-based policy

Based on public concerns about the federal interference in Reading First, Education Week made an open-records request to examine correspondence and documentation in the grant-approval process. They found federal interference in the process. Later, Robert Slavin, the co-founder of the education program *Success for All*, helped launch a federal investigation about the management of Reading First (Manzo, Education Week, March 6, 2007). The federal Office of the Inspector General investigated the issue and then published several reports pointing out flaws in the implementation of scientifically based reading research (Office of Inspector General, 2006; 2007). First, the final report found that federal officials did not screen contractors for potential bias and conflict of interest. It also found that grant reviewers had significant professional connections to a teaching methodology that required the use of specific reading programs, in particular the Direct Instruction model and its Reading Mastery program (Office of Inspector General, 2006, 17). Reading First contractors received royalties from private-sector textbook companies. This private-sector involvement biased Reading First consultants towards recommending literacy products to which they were financially tied. Michael Grunwald, from the Washington Post, characterized Reading First as a "pilot project for untested programs with friends in high places" (Grunwald, Washington Post, October 1, 2006). Reid Lyon from NICHD stated that the Reading First contractors were "actively working to undermine the NRP [National Reading Panel] Report and the RF [Reading First] initiatives" (Office of Inspector General, 2006, 18). Commercial bias guided the implementation of the Reading First initiative.

The Inspector General's investigators found the USDOE had influenced certain states' selection of reading programs (Office of Inspector General, 2006, 2). Similarly, a report

[21] 96th United States Congress (October 17, 1979). Department of Education Organization Act (DEOA).

from the U.S. Government Accountability Office also found that state officials reported receiving suggestions from federal education officials or contractors to adopt or eliminate reading programs (U.S. Government Accountability Office, 2007). States did not always understand the monitoring procedures and the federal expectations. Kentucky and Georgia officials complained that a federal consultant had suggested they adopt a list of core reading programs to improve their chances of getting the Reading First grant (Education Week, 2005). The federal education department held seminars for state officials to help them understand the grant requirements. Three Reading Leadership Academies included panel discussions on reading research, where the majority of panelists represented Direct Instruction (Office of Inspector General, 2007, 8). Audience comments referred to the Academies as a "sales job," a "sales pitch," and a showcase for Direct Instruction (ibid., 9). Textbook companies who perceived themselves as not having federal approval complained that they were losing business because of the states' misconception (Manzo, Education Week, March 6, 2007). According to the Inspector General, many states were under the perception that the Education Department had an approved list of commercial texts for implementing Reading First.

The state of Michigan was the first state approved for Reading First funds. Michigan had proposed to adopt the five best-selling textbooks on the market (Michigan Department of Education, July 1, 2002). They used the University of Oregon's "Consumers Guide for Evaluating a Core Reading Program Grades K–3," which the Institute for the Development of Educational Achievement had developed. The main focus was on the five pillars of the National Reading Panel's report: phonological awareness, phonics, fluency, vocabulary, and comprehension (ibid., 36). Similarly, a Reading First guidance document from the U.S. Department of Education explicitly stated that Reading First programs sought to "embed the [five] essential components of reading instruction into all elements of the primary, mainstream K–3 teaching structures of each State" (U.S. Department of Education, 2002, 2). Federal officials recommended Michigan's list to other states to use in their applications (Grunwald, Washington Post, October 1, 2006). As a result, the majority of the 4,800 Reading First schools had adopted one of five top-selling commercial textbooks, according to the Washington Post (ibid.). There was not

---

Public Law 96-88, 93 Stat. 668.

the hoped-for "dramatic shift" or "new generation" of evidence-based reading instruction, as devised by NCLB (Manzo, Education Week, December 12, 2006). Reading First schools favored products of large commercial publishers, seemingly because those were fastest to formally respond to the NCLB's requirements (Grunwald, Washington Post, October 1, 2006). The previously market-dominant publishing companies were able to expand their market share. The publishers McGraw-Hill, Houghton Mifflin, Hartcourt, and Pearson accounted for 72 percent of the Reading First funding (Manzo, Education Week, December 12, 2006). The textbooks of those companies, however, provided little scientifically based research backing, as defined by NCLB, despite the federal call for evidence-based reading instruction. None of the textbooks had been evaluated using RCTs.

The Inspector General pointed out a general problem of the Reading First initiative: How could the states establish the necessary "scientific base" of research, despite the fact that most reading programs had not demonstrated such a solid scientific base? The approval process worked via the adoption of the five key elements of effective reading instruction, endorsed by the National Reading Panel in 2000 (Sweet, 2004, 25). As long as the publishers emphasized their compliance with those key elements, they could make their case of being evidence based. States then recommended textbooks that demonstrated adherence to the five required essential literacy components (Manzo, Education Week, December 12, 2006). The idea behind Reading First, however, was that States not just check off the alignment with the key program components, but that they consider the scientific evidence of program effectiveness when evaluating reading research (Office of Inspector General, 2007, 17). Ultimately, the Reading First initiative attempted to implement a premature policy. It was as if the Federal Drug Administration had asked doctors in the 1950s to only prescribe RCT-based pharmaceuticals, despite the fact that pharmaceuticals had not yet been tested in that manner. Only a limited number of reading programs themselves had been tested and proven to be effective based on RCTs when the Reading First program began in 2002. The NRP found only one experimental evaluation to demonstrate the effectiveness of McGraw-Hill's Open Court program. Ironically, Success for All later proved to have the strongest record of RCT evidence, but it was excluded from the Reading First initiative (Grunwald, Washington Post, October 1,

2006). Scripted, standardized textbooks like Open Court or Direct Instruction did not incorporate motivational and teacher-driven learning. However, these textbooks had been recommended by the subgroup report of the National Reading Panel, but they were not part of the five pillars of NRP's summary report. In its subgroup report, the Reading Panel had cautioned that many phonics programs tended to "present a fixed sequence of lessons scheduled from the beginning to the end of the school year," although early grade students varied greatly in their skills (NICHD, 2000, 2-97). Elaine M. Garan argued that commercial programs such as Open Court would be incompatible with the NRP recommendations due to their scriptedness (Garan, 2002, 32). Despite its quest for a scientific basis, the RCT movement seemed to have negatively affected the micropolitics of education by promoting commercial textbooks to local school districts without much evidence.

Despite its problematic implementation, did the Reading First initiative increase reading achievement? The federal government set aside 2.5 percent of the Reading First budget for an external evaluation (Gamse & Jacob, 2008). Its main question was: "What is the impact of Reading First on student reading achievement?" Abt Associates was contracted to perform a five-year rigorous, scientifically valid, quantitative study. Because Reading First was a nationwide initiative, an RCT design was not feasible. Abt Associates chose a regression discontinuity design as the strongest quasi-experimental method (cf., glossary). They capitalized on systematic processes that some school districts used to allocate Reading First funds. The evaluators found statistically significant impacts on instructional reading time spent on the five essential components of reading instruction promoted by Reading First. However, they did not find a statistically significant impact on reading comprehension as measured by the Stanford Achievement Test 10 (Gamse & Jacob, 2008). Ultimately, the Reading First legislation did not bring the expected change and turnaround for America's 40 percent of second-graders who were struggling with learning to read.

4. The Federal Priority of RCT designs in evaluating education projects and its controversy

As in medicine, various groups challenged the RCT primacy in education. The No Child Left Behind Law of 2001 and its Reading First initiative fanned a debate about how best to make methodological choices to determine program impact. In particular, the U.S. Department of Education (DOE) proposed a priority of scientifically-based evaluation methods in November 2003 that cited the RCT as the best design for evaluating impact. Representatives of the evaluation community produced several counterstatements, neither of which ultimately influenced the DOE's Final Priority in January of 2005. Only several years and a leadership change later, USDOE began to foster more inclusive thinking in line with previous critics.

The proposed Federal Priority of RCT evaluations

The USDOE published a notice of the Proposed Priority, "Scientifically Based Evaluation Methods," in the Federal Register on November 4, 2003 (U.S. Department of Education, 2003). It was a new funding priority, consistent with the NCLB Act and based on the RCT as the optimal evaluation approach. In the funding allocation process, RCTs received the so-called "competitive preference priority." This meant that when two evaluation designs were of comparable merit, USDOE would select the RCT design over any other design to answer questions of effectiveness. In other words, all other things being equal, USDOE would favor an RCT design over any observational-qualitative design.

The Proposed Priority stated the following:

> Evaluation methods using an experimental design are best for determining project effectiveness. Thus, the project should use an experimental design under which participants—e.g., students, teachers, classrooms, or schools—are randomly assigned to participate in the project activities being evaluated or to a control group that does not participate in the project activities being evaluated (62446).

The department suggested that the most rigorous methods to address the question of project effectiveness should be randomly assigned RCT designs. Moreover, if random assignments were not feasible, then the project might use a quasi-experimental design with carefully matched comparison conditions or a regression discontinuity design. The

priority given to RCT designs also meant that other designs would be ranked as less valuable or invalid for determining effectiveness. The priority stated:

> Proposed evaluation strategies that use neither experimental designs with random assignment nor quasi-experimental designs using a matched comparison group nor regression discontinuity designs will not be considered responsive to the priority when sufficient numbers of participants are available to support these designs (62446).

The one reason for not using experimental and quasi-experimental designs was when the number of observations was too small and such a design was not feasible. In educational programs, however, sufficient numbers of students and teachers were generally available. USDOE thus established a hierarchy of methodological choices when determining the effectiveness of an educational program (cf., TABLE 7).

TABLE 7: Competitive preference priority by the U.S. Department of Education

Design	***Preference***
Experimental design/RCT	Competitive preference priority
Quasi-experimental designs (matched comparison group) and regression discontinuity designs	2nd competitive preference priority
Non-experimental design	No preference

The Proposed Priority was an attempt by USDOE to narrowly define "scientifically-based" research and evaluation in RCT terms, which was already the case in parts of the NCLB legislation. The Proposed Priority did not go well with many different stakeholders. The next section takes a closer look at the evaluation community's response, which in itself was not in agreement.

Criticism of the proposed federal RCT Priority

The Department invited the public to submit comments regarding the Proposed Priority for thirty days until December 3, 2003. During the USDOE response period, the American Evaluation Association's electronic list sparked a discussion on proof of causation. Former AEA president, Michael Scriven, posted the following comment: "Randomized control group trials (RCTs), even when possible, are NOT always superior to other approaches (non-RCTs) in demonstrating causality" (Scriven, Evaltalk, November 12, 2003). Scriven used the example of the causal effect of cigarette smoking

on lung cancer, which researchers had not experimentally demonstrated, but which nobody would question. Non-RCT studies were capable in establishing causal attribution "far beyond reasonable doubt" (Scriven, Evaltalk, November 12, 2003). Scriven argued that RCTs might be the best choice in complex cases, where the researchers would otherwise have to exclude many potential possible causes in order to establish causal attribution beyond reasonable doubt.

AEA's president Richard Krueger informed AEA members that AEA had sent a response to USDOE (Krueger, Evaltalk, November 24, 2003). Several past AEA presidents, such as Michael Scriven and Nick Smith, endorsed the statement. They questioned the Department's privileging of RCTs over other evaluation methods. The AEA statement's major argument was that RCTs were not the only way to determine causality. It referred to epidemiological evidence, such as that found in cancer studies, which were not based on RCTs. The statement emphasized the equal validity of single-subject, observational, quasi-experimental, and RCT designs. The authors did not subscribe to a hierarchy of methodological approaches, which the Federal Priority had proposed.

The AEA statement argued that "the proposed priority manifests fundamental misunderstandings about (1) the types of studies capable of determining causality, (2) the methods capable of achieving scientific rigor, and (3) the types of studies that support policy and program decisions" (American Evaluation Association, November 25, 2003). It found that the RCT priority could lead to "political, ethical, and financial disaster" (ibid.).

In the following discussion, I offer some observations as to how the AEA statement may have judged the limitations of RCTs unfairly. First, I address the suggested methods. Second, I address the issue of laboratory experiments. Third, I discuss the use of mixed methods in general.

First, the AEA statement argued that in order to identify whether, how, and why a program worked, evaluators would not be able to rely solely on RCTs, but would rather need to draw upon a range of social science methods. Such methods include observations,

interviews, case studies, surveys, and other strategies to understand causality (American Evaluation Association, December 3, 2003). Note that three of these suggested strategies— observations, interviews, surveys—were all data collection methods, and not data analysis methods or evaluation designs. Because the RCT is an evaluation design, and not a data collection method, AEA did not suggest alternatives to RCTs, but rather supplementary tools.

Second, the AEA statement argued that RCT approaches would only "examine a limited number of isolated factors that are neither limited nor isolated in natural settings." The authors pointed out the "complex nature of causality." As a result, RCTs would be "less capable of discovering causality than designs sensitive to local culture and conditions" (American Evaluation Association, November 25, 2003). Erroneously, the AEA statement may have re-cast the concept of RCTs as laboratory experiments rather than field experiments that take place by definition in natural settings. Field experiments may still suffer from some level of artificiality due to the research component (e.g., selection process of study population, more frequent measurements and observations of study population). However, Ronald Fisher's original quest for agricultural RCTs was prompted precisely out of a desire to deal with the heterogeneity of treatment units and the complexity of the environment. Fisher showed how experimental researchers did not need to make an exhaustive list of possible uncontrolled causes, but that they were relieved from the anxiety of estimating the magnitude of the innumerable causes (Fisher, 1935, 21). Randomization of treatment and control units would guarantee the validity of the findings if the study sample were large enough (cf., chapter two).

Third, AEA called for the use of mixed methods, which would serve both the discovery and examination of causal effects. RCTs are generally good at discovering causal effects, because they are set up to answer the question of what works in a particular case based on high internal validity. However, examining causal effects, in the sense of interpreting and understanding these effects, might require different approaches. For interpreting effects, qualitative-comparative data analysis and sensitivity to local culture and conditions would be needed. This new emphasis did not mean to replace RCTs, though, but to add on to them.

The AEA counterstatement supporting the Education Department's RCT Priority

It would be misleading to conclude from the AEA statement that practitioners of evaluation were uniformly supportive of this stance. In fact, another group within AEA developed a counter-statement. Its spokesperson, the evaluation theorist Mark Lipsey, posted the statement on the AEA list, calling it "NOT the AEA statement on Scientifically Based Evaluation" (Lipsey, Evaltalk, December 3, 2003). Thomas Cook, Robert Boruch, Peter Rossi, as well as other important theorists who had shaped the emerging field of evaluation, signed the counterstatement. This internal split of the American Evaluation Association had old roots from when the Evaluation Research Society (ERS) and the Evaluation Network merged into AEA in 1986. ERS had followed the experimental evaluation tradition, whereas the Evaluation Network, composed of mostly practitioners in education evaluation, was less strict regarding methods choice.

Lipsey and colleagues did not feel adequately represented by the so-called AEA statement. Thus, they provided counterpoints to the AEA statement and argued that the AEA statement's opposition to the federal notice was "unjustified" and would "represent[] neither the methodological norms in the evaluation field nor the views of the large segment of the AEA membership with significant experience conducting experimental and quasi-experimental evaluations of program effects" (Lipsey, Evaltalk, December 3, 2003). Instead, the AEA statement had been "proffered without prior review and comment by its members" (ibid.).

Lipsey and colleagues argued that RCTs had been "essential to understanding what works, what does not work, and what is harmful among interventions" in various policy disciplines such as medicine and welfare (Lipsey, Evaltalk, December 3, 2003). They clearly supported DOE's proposed notice for prioritizing experimental methods in educational evaluation, and, contrary to the AEA statement, regarded it as a positive development in education policy. The authors referred to the Campbell Collaboration Register, which had recorded nearly 13,000 RCTs in social, psychological, educational, and criminological trials. In contrast, nonrandomized trials had often led to misleading results, which RCTs had rectified.

Mark Lipsey felt that methodological pluralism in the context of educational effectiveness evaluations implied "hostility of the educational research culture to the scientific epistemology on which experimental and quasi-experimental research methods are based on" (Lipsey, Evaltalk, December 11, 2003). In another message, Lipsey found that the "practical effects of AEA's pronouncements in the policy arena" had an undesirable effect: "little likelihood of any actual influence on the final version of the USDOE Office for Innovation and Improvement's review criteria" and "undermining AEA's credibility with the Dept. of Education" (Lipsey, Evaltalk, December 18, 2003).

The AEA-internal exchanges illustrate that its members were not able to reach an agreement as to whether or not to support the Federal Priority of experimental evaluation. On one side, the AEA board felt that the Federal Priority was too narrow in its definition of scientifically-based evaluation and limited the methodological choices to a degree that could be detrimental to policy decisions, which are purely based on RCT designs. On the other side, a group of AEA members stood in the experimental tradition and felt that the AEA board would promote a false and marginal belief in methodological equality. The back-and-forth debates on electronic lists between advocates and skeptics of experimental evaluations were not resolved; they seemed rather to reinforce an ideological gap between the two groups. As a consequence of this disagreement, some members left AEA altogether (Lipsey, email communication, October 3, 2011).

Let us note, however, that differences between the two camps are slighter than they appear at first blush. Both sides might be able to agree on the statement that methodological pluralism could be beneficial in capturing multiple facets in a program evaluation, especially when expanding beyond the "what works" question. One could argue that both statements "RCTs are superior" and "Multiple methods are equal" could be true, if one carefully specifies the context of those statements. Even if experimental designs might be superior in theory with respect to internal validity, they might not be the best choice in applied policy contexts and in understanding why an intervention's process is working.

The Education Department's response to the criticisms

The U.S. Department of Education published their final notice on Scientifically Based Evaluation Methods over one year after their Proposed Priority (U.S. Department of Education, Federal Register, January 25, 2005). USDOE did not change the proposed notice despite many critics, although it summarized the comments by critics and supporters.

First, the Department found twenty-nine comments in support of the priority for random assignment studies, but quoted one hundred and eighty-three comments opposing the priority, most of which came from AEA members. Their main criticism was that random assignment was not the "only method capable of generating understandings of causality" (U.S. Department of Education, Federal Register, January 25, 2005). Other approaches such as observational and single-subject designs would be equally valid in determining causality. In response to the critics, the USDOE characterized the RCT as the most "defensible method in that it reliably produces an unbiased estimate of effectiveness." In contrast, other methods could be misleading compared with experimental evidence—an idea explicitly addressed in the AEA counterstatement (Lipsey, Evaltalk, December 3, 2003).

Second, the Department disqualified the AEA statement's argument for implementing "designs sensitive to local culture and conditions" in order to capture program effects. In response, USDOE recommended complementary case studies that would collect information on local culture and conditions within an RCT design. Those would provide a "deeper understanding of the conditions that may influence the effectiveness of the intervention" (U.S. Department of Education, Federal Register, January 25, 2005). However, these designs could not stand alone to determine causal effects.

Third, USDOE agreed with the comments that the evaluation question must determine the method. However, they insisted that RCTs were best to answer impact questions. The key difference between USDOE and its critics was the granularity of the evaluation question. Whereas USDOE stated that any impact question would prioritize RCT designs, critics argued that causal questions did not automatically imply an RCT approach. The

methodological choice would depend on additional factors. For example, AEA had argued that interventions in complex environments would not favor an RCT approach, but a more qualitative-observational case-study approach.

In sum, the Department did not respond to the cautions and criticisms of a RCT hierarchy of methods, which were brought forward by professional organizations, including AEA. USDOE concluded that RCTs would: a) provide an ideal form of rigorous impact evaluation, b) be able to produce "valid and reliable data," and c) be the only method able to identify a program's causal effect (U.S. Department of Education, Federal Register, January 25, 2005). In this framework, USDOE established a preference to funding RCT designs for evaluating impact. The critics brought forward important arguments for expanding the RCT primacy, which included the need for opening a program's "black box" and investigating the questions of why, how, and in which context a program is effective. As I discuss in the next section, USDOE indirectly responded in incorporating some of these arguments in their statements about methodological choice just a few years later.

Broadening of evidence-based evaluation methodologies

Ultimately, Mark Lipsey was right in predicting that the internal debates within AEA did not produce a strong counterweight to DOE's proposed experimental priority (Lipsey, Evaltalk, December 11, 2003). However, a few years after this incident, USDOE seemed to expand their definition of scientifically based evaluation methods and included quantitative single-case study designs for establishing causal evidence.

A leadership change in the DOE's Institute of Education Sciences made it possible to slightly broaden DOE's definition of scientifically based evaluation methods. In June 2010, the What Works Clearinghouse (WWC) published the "single-case design technical documentation" (Kratochwill, Hitchcock, Horner, et al., June 2010). WWC specifically attempted to "expand the pool of scientific evidence available for review" (ibid., 2). The panel characterized single-case studies (SCS) as adaptations of interrupted time-series designs that could provide "a rigorous experimental evaluation of intervention effects." SCSs were considered experimental because cases could serve as their own

control group. For example, evaluators would repeatedly measure the cases' outcome variables at different points in time. The WWC expanded the concept of experimentation to within-case control designs. This change was a significant shift in experimental thinking, because subjects could become their own control group—an idea excluded in DOE's original priority. To be scientifically based, a SCS would need to fulfill certain quantitative criteria, such as systematic manipulation of the independent variable; systematic measurement of outcome variables over time with inter-assessor agreement; and at least three phase repetitions with three data points each. Although the SCS guidelines advocated for quantitative approaches only, the document was an important step in moving towards a more inclusive approach of causal evaluation. The evaluation theorist Michael Scriven commented that the SCS guidelines finally brought single-case studies doing interrupted time series designs (ITS) the deserved classification on the evidence scale (Scriven, Evaltalk, May 1, 2011). Scriven also pointed out that there were "half a dozen other designs with credentials as good as ITS in suitable circumstances that have cost or ethical challenges in those circumstances, so, even if RCT is FEASIBLE, it may be a stupid waste of resources to use it" (emphasis by Scriven; Scriven, Evaltalk, May 1, 2011). The "half dozen other designs" referred to more qualitative-interpretive data analysis tools, which historical and forensic sciences had frequently used when establishing evidence of causal connections in the past.

John Easton, Director of the Institute of Education Sciences (IES) since 2009, emphasized the relevance and generalizability of education research and evaluation (Easton, speech, March 28, 2011). Easton called for education research "to move beyond trying to discover 'what works' to learning about why, when, where, for whom and under what conditions" (Easton, speech, March 28, 2011, 3; cf. also 11). Easton's voice resembled some of the critics' arguments like the AEA statement to the 2003 Federal Priority. Rather than focusing on the "what works" questions, he demanded evaluation approaches capable of answering contextual questions. IES's new chord was "to work more collaboratively with practitioners and policy makers and build partnerships that engender relevant, useful research" (ibid., 4).

Furthermore, on November 1, 2010, the National Board of Education Sciences approved a new set of research priorities, especially "to understand causal linkages to the greatest extent possible by conducting or sponsoring rigorous studies that support such inferences" (Easton, speech, March 28, 2011, 5). Understanding causal relations would move beyond just identifying the causal effects, which RCTs are suited for.

Easton expressed the importance of moving away from "simple black box or silver bullet approaches" towards investigating the mechanisms of school improvement (Easton, speech, March 28, 2011, 11). And finally, Easton summarized his quest:

> If we are asking our research to answer more complex questions, it also means we must expand our repertoire of rigorous methodologies. Moving forward, I believe IES [Institute of Education Sciences] should investigate mechanisms and moderators using data from randomized trials; allow for the analysis and use of quasi- and non-experimental evidence for studying schooling processes and context; and measure program implementation, fidelity and sustainability. We can apply the same effort to building rigor into these methods as we have to RCTs.

Although Easton does not suggest non-experimental methods for establishing impact, he advocated for the use of non-experimental evidence for answering the questions of how schooling works. Answering "how it works" questions are policy relevant because they allow for applying evaluation findings to other contexts. This broadening of rigorous methodologies in education is an encouraging direction in the policy arena as it promises the production of more policy-relevant evaluations. Frustrated by unscientific evidence in education research, the National Reading Panel and the U.S. Department of Education had made the RCT the apex in the methodological tool box. Reading First tried to implement a scientifically based reading policy nationwide. However, RCTs did not provide much evidence (cf., the WWC struggle for finding effective interventions), and textbook companies found ways to circumvent the experimental demand. Professional evaluators such as from the American Evaluation Association argued that non-experimental research can produce evidence that is not only scientific, but also policy relevant. The WWC took the first step by setting the standards for the single case study design in 2010. According to Easton, this venturing into new methodological approaches would continue and arrive at more inclusive methodological choices. Ideally, this process would build bridges between RCT critics and RCT advocates. The future will reveal

whether methodological rigor and policy relevance could further join forces to inform program evaluation designs, and make and implement better education policies.

Concluding remarks

The STAR experiment, the National Reading Panel, and the Reading First implementation illustrated several challenges in pursuing policies based on experimental evidence. They also demonstrated how the implementation of RCT-guided standards affected micropolitics in an unexpected—and mostly negative—way.

With regard to the STAR experiment, the findings were internally valid for the tested population. However, due to self-selection, exclusion criteria, and attrition, findings were less generalizable than anticipated. STAR did not answer the question of how and why class-size reductions were effective. When the state of California tried to implement a similar class-size reduction policy, they faced a shortage of qualified teachers. California was unable to replicate the STAR miracle, and the state was left with unintended negative consequences, such as under-funded schools and an under-qualified teaching staff. Extrapolating findings from pilot settings, such as STAR, may lead to deficient policy solutions. The temporal-spatial and contextual constraints of an RCT often do not allow for generalizations to different geographical and temporal settings.

In the example of the National Reading Panel, the application of experimental evidence to a particular policy area proved difficult. The process of selecting topics seemed less rigorous than the establishment of evidence-based standards. The choice of topics, however, would already influence the nature of the findings. One problem was that the RCTs were unevenly spread across the range of reading interventions. The panel members then chose to review studies in areas where sufficient experimental evidence existed. The NRP did not find sufficient experimental evidence to determine best practices in several areas of reading research, as might occur in more holistic reading interventions. Many reading approaches, thus, were deemed unproven. Although the NRP attempted to move away from politics toward a scientific consensus, the unevenness of the findings led to the questioning of its academic integrity. As a result, the debates over effective reading interventions did not get settled; rather, they intensified.

Similarly, the quest for policies based on science did not pay off in the way planned in Reading First. Its goal was to exclude potential political bias by endorsing the RCT as unbiased methodology to select effective reading programs. The shortage of RCT evaluations led textbook companies to align their curricula with the five NRP pillars of literacy, thereby circumventing the original quest for rigorous evaluation. Ultimately, commercial biases rather than a scientific quest for truth dominated the implementation of Reading First.

The section on the Federal Priority provided an example of the debate over the RCT primacy between the U.S. Department of Education and professional organizations. DOE's proposed 2003 Federal Priority considered the RCT as the best design for evaluating impact. Although the AEA and other professional organizations criticized the planned priority and suggested alternative designs, USDOE did not change its priority. Several years later, however, USDOE became more open to alternative designs due to the promise of greater policy relevance.

In sum, the RCT model in U.S. education policy faced major challenges. First, a shortage of RCT evaluations in many areas such as motivational learning led to the exclusion of these areas in policymaking. Second, policy makers were overly reliant on already-existing RCT findings, such as class-size reduction or phonics instruction, and used their findings beyond the original scope. Because education interventions are less discrete than the administering of drug treatments, more knowledge is needed about their context and underlying processes. RCTs simply do not provide this knowledge when applying findings outside the experimental population.

Ultimately, the use of RCTs in the field of education was less successful than in medicine, due in part to the specifics and complexities of the schooling context. The multitude of factors that influence programs and their contexts make RCT findings much less valuable in education than in medicine. Even if an RCT yields positive findings (such as in STAR), little knowledge is gained. Policy makers need information about the underlying processes of why an education intervention works and how to adapt it to

different contexts. John Easton from the Institute of Education Sciences advocated for such studies, often non-experimental in nature, that would explain such processes. A look at the newest developments in medicine, such as Personalized Medicine, could guide the researcher away from RCT-based average effects on the average student toward more context-specific findings.

CHAPTER 4: THE MEDICAL RCT MODEL IN INTERNATIONAL DEVELOPMENT

As in education policy, there has been a renewed interest in the role of RCTs in international development. Advocates of RCTs often refer to Progresa, Mexico's Conditional Cash Transfer Program, as a prime example of a successful RCT in development. In the first section of this chapter, I analyze the Progresa trial and argue that, despite a decade of RCT studies on Progresa, many important policy questions remain unanswered. A major outstanding issue, for example, is whether conditionality—i.e., program participants receiving money only when certain conditions are met—is indeed needed to obtain the desired impact.

The second section starts by examining the Logical Framework Approach (LFA), which USAID adopted in 1971 and exported to other bilateral and multilateral institutions. I show how the LFA established the need for capturing development impact via a results chain, despite failing to suggest the best means for doing so. I argue that this indeterminacy may have later triggered the debate over the role of RCTs in measuring development impact. In reference to RCT evaluations used in drug testing, the Center for Global Development (CGD) published a controversial report in 2006 calling for rigorous impact evaluations.

In the third section, I analyze the RCT debate illustrated by the Network of Network on Impact Evaluation (NONIE), which was a multilateral response to the CGD report. NONIE attempted to establish policy-relevant methodological guidelines for impact evaluations. I discuss how and why NONIE members could not agree on the role of RCTs or how they should figure in a decision-tree of methodological choices. Although the final draft of the document integrated various views, not all members accepted the document. Furthermore, the European Evaluation Society vocalized criticism against the final NONIE draft (cf., section 4). Finally, the U.S. Agency for International Development (USAID) felt compelled in 2011 to publish an evaluation policy, which celebrated the diversity of methods for evaluation in general. At the same time, USAID

gave special value to RCTs in impact evaluation. In sum, the influence of the RCT debate is apparent in official documents, which show great caution with respect to the role of the RCT in impact evaluation. I demonstrate that the methodological pendulum, which had been swinging towards the RCT side, is moving back to more middle ground. This trend is arguably due to the fact that RCT findings have not been sufficiently relevant for policy makers and need to be supplemented or even supplanted by other methodologies.

1. The RCT of Mexico's conditional cash-transfer programs and some unanswered questions

Advocates of the RCT model often refer to successful examples of RCTs in order to demonstrate the model's superiority. The 1997–1999 RCT evaluation of Mexico's Progresa, a Conditional Cash Transfer (CCT) program, functions as a poster child of a successful RCT for development evaluation—much like the Tennessee Teacher Student Achievement Ratio (STAR) trial functions for RCTs in U.S. education (cf., chapter three). Economists Abhijit Banerjee and Esther Duflo, for example, cited Progresa as one of the "first demonstrations of the persuasive power of a successful randomized experiment" (Banerjee & Duflo, 2011, 79). The Evaluation Gap Working Group of the Center for Global Development considered Progresa a good example of a high-quality RCT evaluation (Evaluation Gap Working Group, 2006, 3; 18; 25). They pointed out that the RCT evaluation was able to put to rest serious concerns from political opponents who argued that giving funds to poor mothers might increase their vulnerability to domestic abuse. (Evaluation Gap Working Group, 2006, 23). Furthermore, Thomas Vinod, director of the Independent Evaluation Group at the World Bank, stated that the Progresa experiment provided evidence "when these [types of] programs were being dismissed by development practitioners" and helped to "depoliticize" decisions to some extent (Vinod, presentation, February 20, 2008).

Given its role as a poster child, Progresa warrants a closer look. Researchers must consider what lessons should be reviewed in the current debate over the value of the RCT model in development. I conclude that despite Progresa's prominence, many questions are still unanswered; in particular it is unclear how representative the findings are and

what the necessary components of the program are. In fact, it is still unclear whether conditionality—Progresa's core component—is necessary for program success.

Background of Progresa's RCT evaluation

The RCT evaluation of Progresa stands in the tradition of medical trials, having expanded from developed to developing countries early on. The U.K. Medical Research Council, for example, not only conducted the Streptomycin trial in the United Kingdom, but also implemented subsequent Tuberculosis trials in India through testing the necessity of bed rest (Dawson, 1966;Valier 2008, 659). Albert Sabin chose the former Soviet Union and other parts of Eastern Europe as sites for his Poliomyelitis trials starting in 1955 (Oshinsky, 2005, 252). Since the 1960s, researchers have tested the effects of nutritional intervention on the stunting of individual's growth in developing countries (Behrman, 2009, 1372). Following the tradition of clinical trials in medicine, such nutrition interventions were typically accompanied by an RCT. One major difference to medical trials was that randomization was often not feasible at the individual level due to contamination. Instead, the community became the unit of randomization. The nutritional intervention in Guatemalan villages (1969–1977), for instance, demonstrated the use of randomization at the community level to evaluate program effectiveness (Scrimshaw, 2010; Martorell, 1995a; Matorell at al., 1996; Matorell, Habicht, & Rivera, 1996). RCTs were later expanded to anti-poverty programs in developing countries—Progresa being one of the first.

Progresa was initiated in 1997 by the Mexican government as a pilot program that provided conditional cash transfers to improve educational attainment, health, and nutrition for poor rural households (Progresa, 1999; Progresa, 2000). The ultimate goal of the program was to reduce intergenerational poverty. Prior to Progresa, food subsidies had been the primary means of targeting poverty.

Through Progresa, the Mexican government initiated a pilot project with two features that broke with traditional social programs: first, conditionality of monetary assistance, and, second, the built-in RCT evaluation. First, the assistance consisted of direct monetary transfers instead of food supplies, and the beneficiaries only received the money on the

condition that they assumed responsibilities for a series of tasks. For example, women were required to get free regular medical check-ups during pregnancy and lactation. They were also required to send their school-age children to school. If she fulfilled the conditions, a mother of a ninth-grade daughter would receive 255 pesos a month, which equaled approximately sixty-seven percent of what her daughter would have earned if she worked instead of attending school and 20 percent of a household's average monthly income. Active citizenship and comprehensiveness were key features of the program's design. Progresa's architect, Santiago Levy, stressed that "shared responsibility and respect" in a democratic society link the cash benefits to concrete actions of household members (Levy & Rodríguez, 2004). Families would take direct action to improve their own nutrition, education, and health (Behrman, 2007). According to Lucy Luccisano, Progresa represented the concept of "government through freedom"—that is, a shift from paternalistic governing to governing through the active and responsible choice of individual citizens (Luccisano, 2004)., A centralized distribution of funds to local communities encouraged local agency and promised a reduction in the opportunities for corruption. The use of local service providers was a desirable concept for policy makers in the context of decentralization thinking.

The second feature that set Progresa apart from other social programs was the RCT evaluation for testing the effectiveness of the novel design. In a 2009 keynote address, former President Ernesto Zedillo stated he and others had designed Progresa using scientific evaluations and measurements already built in from the outset (Zedillo, speech, April 17, 2009). At the time of the design, a major concern was what would happen to the program during the change of Mexico's political leadership. In 1997, the program designer gathered political buy-in by obtaining the state governors' approval before Progresa had even started. The Mexican Government contracted with International Food Policy Research Institute (IFPRI) to conduct an independent, external, and quantitative evaluation (International Food Policy Research Institute, 1999). To make randomization feasible and reduce possible contamination effects, communities rather than individual households were randomly assigned to treatment and control groups using a pipeline approach with a 20-months lag. The data from the control group provided a natural benchmark against which to judge how the treatment group would have fared without

Progresa. IFPRI surveyed 24,000 households in 506 villages on five occasions between 1997 and 1999. Later, the Mexican Instituto de Nutrición y Salud Pública performed further evaluations, conducting 160 qualitative focus group interviews in 88 sites (Adato, Nov 24, 2008). President Zedillo reported that they were able to attract excellent researchers worldwide (Zedillo, speech, April 17, 2009). The Mexican government shared Progresa data with other scholars. As a result, Progresa became one of the most studied programs in the developing world (Lustig, 2011, 8).

Results of the Progresa RCT

The evaluation results came in between presidents and were strategically primed to influence program survival. The program survived the transition from the seventy-year ruling of the socialist Partido Revolucionario Institucional (PRI) to Vincente Fox's center-right Partido Acción Nacional (PAN). According to Alan Krueger, the concept of conditional cash transfer survived the change of party leadership, possibly due to its rigorous evaluation (Krueger, New York Times, May 2, 2002). The attribution to Progresa's survival is hard to make, however, because it consisted of several new features apart from the RCT evaluation. It also represented a new way of social programming.

The evaluation results of Progresa showed a statistically significant increase in school attendance in Progresa villages compared to control sites. The randomized evaluation found an eight percent increase of girls from 67 to 75 percent and a four percent increase of boys from 73 to 77 percent to attend secondary schools (Schulz, 2004). However, the evaluators expected a higher increase: "The inelasticity of the demand for schooling still poses a puzzle" (Schulz, 2004, 222). The RCT could not pinpoint the reasons for this modest increase.

Despite rather small findings, Progresa became a model for social interventions worldwide. As such, academics, multilateral organizations, policy makers, and the media celebrated Mexico's Progresa program as a model of successful antipoverty programming. According to the economist Nora Lustig, the factors that accounted for Progresa's success were that well-trained scholars transformed into influential practitioners who then played a fundamental role in promoting the new conceptual

approach of poverty reduction (Lustig, 2011, 2). These scholar-practitioners ensured the technical soundness and effectiveness of the program's design, incorporating rigorous impact evaluations in the program's design and ultimately persuading politicians to implement and keep the program in place.

The expansion of conditional cash transfer programs

Policy makers in the development arena regard the conditional distribution of cash as one preferred method in social development, as opposed to unconditional cash transfers such as assistance payments or pensions. The 2009 World Bank synthetic review on conditional cash transfers characterized CCT programs as "modernization of social assistance" (Fiszbein & Schady, 2009, 100). Instead of "pure handouts," political decision makers favorably regard conditionalities as "co-responsibilities" in a social contract with the poor (Fiszbein & Schady, 2009, 10). The introduction of conditions would increase the overall budget available for redistribution due to acceptance across the political spectrum (Fiszbein & Schady, 2009, 60).

Based on such positive perception, CCT programs exponentially expanded within the next decade. In 1997, three CCT programs existed in three countries: Mexico, Brazil, and Bangladesh. In 2008, 28 CCT programs were in place across the globe, such as in Nicaragua, Jamaica, Chile, Malawi, Zambia, and Indonesia (Fiszbein & Schady, 2009). In the United States, Opportunity NYC in New York, and the Capital Gains Program in Washington, DC have been experimenting with conditional cash incentives to high school students and their parents for attending school, involvement in parent-teacher meetings, and passing standardized tests. The New York Mayor's office explicitly modeled Opportunity NYC after Progresa and traveled to Mexico to learn about the program in action. RCTs accompanied the U.S. programs (Fiszbein & Schady, 2009, 144).

The culture of evaluation around CCT programs has been strong beyond traditional practice in social policy areas (Fiszbein & Schady, 2009, 7; 94). Similar to Progresa, CCT programs most often include a rigorous experimental or quasi-experimental evaluation, using "credible counterfactuals" (Fiszbein & Schady, 2009, 7). Several CCT

programs used government-external—and even country-external—evaluators. In the case of Nicaragua, the very IFPRI that had conducted the Progresa evaluation a few years earlier conducted an RCT (Maluccio & Flores, 2005; Maluccio, 2005). IFPRI also included an ethnographic case study in six Nicaraguan sites. Fieldworkers lived with households in the community for three to five months, observed those households, and interviewed all household members (Adato, Nov 24, 2008). Despite the fact that survey data suggested that mothers gave iron supplements to their children, ethnographers found that parents had not administered the iron (Adato, Nov 24, 2008).

Discussion: Conditional cash transfers programs and their results

Despite the fact that CCTs have been evaluated so widely and extensively—often utilizing RCTs—there remain major open questions for policy makers. These include questions about excluded populations, long-term impact, necessary components (especially regarding conditionality), unintended consequences, and transferability of findings.

The 2009 World Bank synthetic review called CCTs an "effective way of redistributing income to the poor" (Fiszbein & Schady, 2009, 30). However, despite the "accumulating evidence of positive impacts" (Fiszbein & Schady, 2009, 41), the report found that many details were not yet known. For example, the majority of evidence stemmed from middle-income countries in Latin America (Fiszbein & Schady, 2009). The CCT adaptability to diverse country settings was still understudied.

Questions about excluded population: In order to be implemented, CCTs require certain structures in place. Such structures include basic national and local coordinating structures for implementing the distribution of funds, for the monitoring of conditions met, and for the provision of the services offered in health and education. To meet the conditionality, poor households needed easy access to health clinics and schools, which might not be the case in remote areas of developing countries. The Progresa program excluded communities from participating when they did not provide health or education services. This criterion meant that all poor households who lived in communities without the minimum service capacity were excluded from the CCT program (Fiszbein

& Schady, 2009, 75). In Columbia, 15 percent of communities were automatically excluded because of infrastructure deficits. This problem is similar to the "general run of patients" argument, where only hospitalized patients participated in a certain medical trial (cf., chapter 2). Just as it was unclear how other patients outside of the hospital setting would fare in the medical program, it is unclear in the CCT case how households in communities without infrastructure would fare on the program.

Questions about long-term impact: One of the major goals of CCT programs is to have an intergenerational effect. Ideally, policy makers should be able to determine whether CCTs affect years and quality of schooling completed by adults. By design, Progresa's pilot results focused on short-term effects as the pipeline design introduced conditional cash transfers in the control communities twenty months later. So far, one long-term evaluation of Progresa showed that students finished an average of one-fifth of a year of more schooling (Behrman, Parker, & Todd, 2005). Researchers did not find an impact on learning outcomes. In general, CCT evaluations showed increased household consumption and increased use of education and health services. However, most RCTs did not study long-term impact. If they did, evaluators found only modest effects on final outcomes in years of schooling completed, cognitive development, health condition, learning outcomes of children, as well as the impact on long-run autonomous family incomes and poverty reduction (Fiszbein & Schady, 2009, 21; 96; 127).

Questions about necessary components: CCT programs are comprehensive, multidimensional programs, with multiple conditionalities, short-term objectives, and long-term goals. They are packages that have several components, as illustrated in Table 8. Many of these components might be highly context specific and dependent on various factors.

It is unclear which program components are important in achieving the particular outcomes (Gaarder, presentation, May 4, 2010). Which households should be selected? How large should the payment be in proportion to the household income? Should school attendance be monitored? These and other questions need to be asked when designing a CCT program. In principle, the particular results (e.g., increased school grades) could

stem from the income effects associated with the transfers (e.g., the child no longer needs to work and could dedicate time to homework), or they could stem from the condition of school attendance, or both. The issue of why the program is effective is important for program design. Understanding why would help determine factors such as optimal size of transfer, what conditions to use, how to monitor them, and how to penalize non-compliant beneficiaries. Social, cultural, and economic factors might affect and be affected by the CCT program. Program impacts are typically mediated by social processes in households and communities of individuals with their own culture, beliefs, experiences and interests (Adato, Nov 24, 2008). Understanding and influencing these social factors could lead to increased impact.

TABLE 8: Components of conditional cash transfer programs

Component	*Examples*
Eligibility based on selection criteria	e.g., threshold of household income
Community characteristics	e.g., required minimum infrastructure of service provision in communities
Payee	e.g., male or female head of household; teenage child
Size of payment	e.g., as a proportion of household income
Form of payment	e.g., cash in form of wire transactions, debit cards, bank account transfers;
Timing of payment	e.g., monthly, bi-monthly, quarterly
Conditions	e.g., school attendance, passing grades, health check-ups, child immunizations, health education
Timely monitoring of compliance	e.g., monthly to yearly or none
Enforcement and sanctions	e.g., no monitoring, warnings of social workers before terminations, immediate termination

Is the conditionality necessary? Policy theorist Carolyn Heinrich pointed out that most impact evaluations have not answered the conditionality question (Heinrich, presentation, April 18, 2009). Are conditions needed? The comparison group typically consists of a no-treatment condition, as in the case of Progresa. Only preliminary research exists that compares the treatment communities receiving conditional cash with control communities receiving unconditional cash transfers. Small-scale experiments in Malawi and Morocco have thus far incorporated comparative treatment arms with conditional and unconditional transfers (Banerjee & Duflo, 2011, 80). In both instances, researchers

found that the conditional program did not significantly outperform the unconditional program, though it did much better than the no-treatment group (Benhassine, Devoto, Duflo, Dupas, & Pouliquen, 2010; Banerjee & Duflo, 2011, 283). Ideally, more of these designs would be needed to tease apart the effects of the transfer from the effects of the conditions.

Further preliminary findings indicated that conditionality was not required for a cash transfer program to have an impact on child well-being. Ecuador's Bono de Desearollo Humano program originally included conditions in the design, and a public campaign had emphasized the human capital goals of the program (Fiszbein & Schady, 2009, 156). The program dropped those conditions in the implementation process (Paxson & Schady, 2007). In the RCT conducted on the program, Paxson and Schady found that the unconditional cash transfer program raised household consumption levels; furthermore, physical, cognitive, and socio-emotional development of preschool children were higher than in the control group (Paxson & Schady, 2007). No data were collected for school-age children.

Questions about unintended consequences: Conditional cash transfers are by design only intermediate means of poverty reduction. Giving cash to people would not be an ultimate policy solution to social programs, especially when one regards market-driven economic growth as an ideal means of poverty reduction. Heinrich raised the question as to whether CCT programs encourage individuals and households to raise their own productivity and income, which ultimately would make government-provided cash assistance superfluous (Heinrich, presentation, April 18, 2009). A further issue is what other unanticipated effects the CCT programs might have on the household. For example, given the increase in household income from the CCT, labor force participation might drop, distribution of labor within families could be reconfigured, and in general, household economies might be significantly altered. Finally what are other spillover effects such as socioeconomic status within the community? (Heinrich, presentation, April 18, 2009; Fiszbein & Schady, 2009, 96). As of yet, none of these questions have been satisfactorily answered.

Questions about transferability of findings: Even if the findings in Mexico were promising, that does not mean that the program would work in other countries. In Mexico, where almost universal primary-school attendance had already existed in the 1990s, the policy focus was on participation in secondary education. In the United States, the CCTs in Washington, DC and New York focused on learning outcomes rather than school attendance, because secondary school participation was already high (Fiszbein & Schady, 2009, 179). Consequently, CCTs might not be the right choice for income distribution and poverty reduction in certain contexts. CCTs are only one option within the range of social assistance programs. They might also not be a stand-alone program, but would benefit in tandem with other programs.

Even after 10 years of experimentally and rigorously evaluating CCT programs, the World Bank review concluded: "We cannot tell at this time whether the current wave of CCT programs will be successful in unleashing a sustainable transformation" (Fiszbein & Schady, 2009, 203). In order to find out whether CCTs are more than just promising policy solutions and not a mere fad, further evaluations are necessary. Evaluators will need to open the so-called "black box" of the process underlying how CCTs produce the intended results (Bourguignon & Sundberg, 2007).

2. The Evaluation Gap Working Group's report: call for rigor in impact evaluation methodology

In 2006, the Center for Global Development published a report calling for rigorous impact evaluations, which triggered debates about the role of RCTs in development evaluation. The group did not reach a conclusion as to whether the RCT should be the best method for impact evaluations.

Historical background: Aid Accountability and the Logical Framework Approach

Foreign aid assistance started to grow after the Second World War as did the congressional call for accountability. In 1971, Leon Rosenberg, a private contractor, developed the LFA for the U.S. Agency for International Development (USAID) to model an intervention "logic" from program inputs to program impact.

The LFA consisted of a 4 x 4 matrix (cf., Figure 2). The first column described the program activity, which included: impact, outcomes, outputs, and inputs. Other elements were indicators (i.e., measures of the program activity), means of verification (i.e., procedures to collect information about indicators), and assumptions (i.e., underlying assumptions about the link between the LFA elements in the first column).

FIGURE 2: Illustration of a Logical Framework Approach

Project Structure	**Indicators**	**Means of Verification**	**Assumptions**
Impact or goal	How the achievement of the impact will be measured	Sources of information on the impact indicators	Assumptions concerning the outcome to impact linkage
Outcome or purpose	How the achievement of the outcome will be measured	Sources of information on the outcome indicators	Assumptions concerning the output to outcome linkage
Output	How the output will be measured	Sources of information on the output indicators	Assumptions concerning the input to output linkage
Input	How the input will be measured	Sources of information on the input indicators	Assumption concerning the input linkage

Adapted from USAID (2005)

Consider the following example of conditional cash transfer: Inputs (financial and human resources; e.g., cash and teachers) were expected to produce certain outputs (e.g., students attend school), which were expected to lead to certain outcomes (e.g., students' successful graduation), and which in turn would generate long-term impacts (e.g., finding employment, increased consumption, better life).

The assumptions column lists all assumptions that would need to hold in order for the impact chain to occur. For example, the schools would need to employ teachers, students need to be present at schools and pay attention. Rosenberg regarded the links among levels of output, outcome, and impact as always hypothetical: "It is a hypothesis that

achieving the results expected at each level will lead to achieving the results expected at the next higher level" (Rosenberg, report, July 24, 1970, 2).

In the original LFA reports to USAID, Leon Rosenberg asked the project staff in the field to formulate their projects:

> "Think about your project as an experiment in economic development. The project has been undertaken because of our conviction that the results will justify the resources provided; however, we want to be explicit about the impact expected of the project and our hypothesis that our inputs will tip the scale to cause that impact" (Rosenberg, report, July 24, 1970, 19).

When Rosenberg used the term "experiment," he did not refer to an RCT, but more of a thought experiment: What would happen if one designed projects a certain way? Rosenberg did not provide insights on how to determine whether the project caused the impact. He only generally suggested testing the hypothesis by generating evidence in support of the hypothesis (Rosenberg, report, July 24, 1970, 20).

In 1987, Richard Solem from USAID reviewed the use of Rosenberg's LFA. Logical frameworks in the USAID tradition became the preferred conceptual approach for most multilateral and bilateral aid institutions to guide the design, implementation, and evaluation of their development interventions (Solem, April 1987, 16pp.). They adapted the system, and there was fluency in the terminological choices.

Solem described the idea behind the LFA as pretending to be "chessmasters"—that is, to be able to see beyond the immediate actions, to project consequences, and ultimately to project impacts (Solem, April 1987, 27). The LFA provided a "common playing field and language" for its multiple players, including project managers and policy makers (Solem, April 1987, 28). Solem found the LFA's "basic power" to be the ability to show causality in the project structure column and dependency on exogenous variables in the assumptions column (Solem, April 1987, 17). He pointed out a tendency to precisely determine inputs and outputs to many intervention efforts, but he saw little clear linkage to specific outcomes and impacts (Solem, April 1987, 24). Furthermore, Solem found that assumptions were often enormously speculative, leading to un-substantiated predictions (Solem, April 1987, 30). The input-output-outcome-impact chain tended to break after

the input-output linkages. Although goal-oriented terminology was already present in international development since the 1970s, methodological means to determine whether a project truly achieved impacts were not yet the central focus.

The Logical Framework Approach (LFA) provided a common framework for communicating about the intended impacts of aid amongst the stakeholders, but no process existed for verifying that the observed consequences were causally connected to the intervention.

Despite bilateral and multilateral efforts to promote evaluation thinking via logical frameworks, internal evaluations often failed to adequately determine long-term impacts—i.e., they could not produce the evidence needed to move from outputs to impact within the LFA. Because impacts could not be adequately measured, donors would never know the economic and social effects of their investments (Leeuw, 2005).

The beginnings: The Evaluation Gap Working Group's quest for evaluating impact

In 2004, with funding from the Gates Foundation and the Hewlett Foundation, the Center for Global Development (CGD) convened a working group to tackle two areas: first, to investigate why rigorous impact evaluations of social development programs were relatively rare; and second, to propose how to stimulate more and better impact evaluations. The working group leaders from CGD were William Savedoff, Ruth Levine, and Nancy Birdsall. The group stood in the tradition of aid critics and maintained the underlying assumption that the impact of past aid had been unclear:

> "Yet after decades in which development agencies have disbursed billions of dollars for social programs, and developing countries and nongovernmental organizations (NGOs) have spent hundreds of billions more, it is deeply disappointing to recognize that we know relatively little about the net impact of most of these social programs" (Evaluation Gap Working Group, 2006, 1).

The group posited that the World Bank and United Nations agencies might have spent billions of dollars on dubious development projects without knowing whether these projects would make a positive impact on people's lives.

In September 2005, the working group leaders presented an initial consultation draft on the CGD's website for comments.[22] In May 2006, the CGD launched the report at the Rockefeller Center in Bellagio, Italy, and invited representatives from international organizations, bilateral donors, private foundations, and developing countries (Center for Global Development, June 12, 2006). This final report had the provocative title "When Will We Ever Learn? Improving Lives Through Impact Evaluation," which called for more and higher-quality impact evaluations. The report found a gap in both the quantity and quality of impact evaluations (Evaluation Gap Working Group, 2006, 10). Regarding the low quantity of impact evaluations, it reported that few incentives existed for donor institutions to produce evaluation reports that would generate valid evidence and gauge program effectiveness; they typically produced monitoring reports. In the LFA scheme, they measured the lowest level, i.e., the link between input and output. The working group, however, was concerned about "impact" (i.e., the final level of the LFA; cf., previous section), and they called for rigorous investigation as to whether a development intervention had reached this final level of the results chain.

These rigorous impact evaluations would fall outside the organization's normal budget and planning cycles (Evaluation Gap Working Group, 2006, 2). Insufficient valid evidence would lead to confusion over the effectiveness of development interventions. In the absence of knowledge, funders tended to operate under the assumption that any development intervention would at least have some positive effect.

Regarding the low quality of impact evaluations, the group found the quality of existing evaluations to be poor. These were concerned mostly with outputs rather than impacts of development programs. Even when evaluations attempted to measure effects and impacts, poor methodological choices would lead to an overestimation of positive effects. Instead, impact evaluations would need to test the "net effect" directly attributable to a specific intervention. The report stated that "no responsible physician would consider prescribing medications without properly evaluating their impact or potential side effects" (Evaluation Gap Working Group, 2006, 3), and therefore clinical trials had become "the standard and integral part of medical care" (Evaluation Gap Working Group, 2006, 15).

[22] Cf., http://www.archive.org (http://www.cgdev.org) for access.

Referencing the medical model of experimental trials indicated the group's preference for RCT evaluations.

Are rigorous impact evaluations limited to RCTs?

The report called for more and more rigorous evaluation approaches that tested whether development interventions had an impact—that is, whether they produced the results they were originally pursuing. The group defined impact evaluations as "studies that measure the impact directly attributable to a specific program or policy, as distinct from other explanatory factors" (Evaluation Gap Working Group, 2006, 10). Impact evaluations gauged the "net impact" of an intervention, which was the change of the condition that the intervention sought to alter, minus all the other factors that simultaneously affected the conditions (Evaluation Gap Working Group, 2006, 2). From a methodological stance, a core concept of investigating the net impact was the "counterfactual," i.e., what would have happened without the program. Therefore, the report posited: "Impact evaluation asks about the difference between what happened with the program and what would have happened without it (referred to as the counterfactual)" (Evaluation Gap Working Group, 2006, 12). An example would be whether children in Progresa villages would stay in school and learn more than they would have without the Progresa program. The net impact would be the gain in yearly attendance or the increase in test scores, based solely on the Progresa program. The evaluator, however, would never be able to encounter the counterfactual. To simulate the counterfactual, the evaluator would need a comparison group:

> Most notably, [impact evaluations] require attention to gathering information from appropriate comparison groups so that valid inferences can be made about the impact of a particular program compared with what would have happened without it or with a different program. This type of data collection must be considered from the start (Evaluation Gap Working Group, 2006, 13).

The comparison groups would simulate the counterfactual situation. Evaluators would need to collect data on the conditions of interest before and after the intervention. From this point, it is only a short leap to recommending the RCT as the ideal evaluation design, because the treatment and the control group characteristics would be equally distributed via random assignment to validly estimate the impact.

In fact, the initial consultation report from September 2005 suggested that RCTs were the authors' preferred method. This initial report, however, explicitly stated that it represented only the views of the working group leaders (Savedoff, Levine & Birdsall, report, September 15, 2005, i). The authors stated in a footnote that a major area of debate was whether the working group should indicate a preference for RCTs:

> The extent to which the "Club" should favor random assignment studies was discussed extensively by the members of the Evaluation Gap Working Group. The consensus was to recognize and support random assignment studies, however the group did not reach a specific conclusion about how much of the club's funding should be earmarked for random assignment studies (Savedoff et al., 30).

According to the authors, because the working group's mandate was merely concerned with impact evaluations, focused attention on RCTs was warranted (Savedoff et al., report, September 15, 2005, appendix 10). The authors suggested that: "Because studies with randomized assignment face the largest obstacles relative to their promise in knowledge building, more than half of the Club's funds should be earmarked to support studies with randomized designs." (Savedoff et al., vii).

The second draft did not directly state a preference for RCTs. However, the focus on the concepts of "net effect," "counterfactual," and "comparison groups" indirectly leaned toward the RCT as the preferred approach. As mentioned, the group anchored their arguments in the success of rigorous clinical trials in medicine. They stated that the reason why clinical trials of medications have become a standard and integral part of medical care was to counter the risk of wasting public resources or harming participants (Evaluation Gap Working Group, 2006, 15). Although the history of medical trials did not indicate a primary concern for wasting public funds—but rather the patient's money—aspects of harm and safety were driving forces in the requirement of well-controlled clinical trials before a medication's marketing approval (cf., Drug Efficacy Amendment of 1962 in chapter two). The working group's direct reference to the medical field indicated a preference for RCTs.

They recommended that researchers "choose the best method" for collecting and analyzing data and drawing valid inferences. In their view, "it is usually worth asking whether a random-assignment approach—that is, randomly choosing which individuals,

families, or communities will be offered a program and which will not—is appropriate and feasible" (Evaluation Gap Working Group, 2006, 15). If RCTs were not feasible, evaluators could apply other evaluation approaches, including well-controlled before-and-after studies, interrupted time-series studies, and matched comparison studies (ibid., 15). In the tradition of the Campbell Collaboration and the Cochrane Review, the working group indirectly established a hierarchy of methodological approaches to generate evidence, which the working group explicitly cited as sources for finding evidence-based health interventions (Evaluation Gap Working Group, 2006, 17; 30).

Overall, the working group was foremost concerned with methodological and technical issues in development evaluation, in particular the issue of "internal validity" (Evaluation Gap Working Group, 2006, 17). They cited methodological shortcomings and gave negative examples of flawed methods, such as before-and-after assessments and evaluations with incompatible comparison groups. In the appendix, they quote D. Levine, stating that because RCTs were "more convincing than other research methods, most evaluations should involve randomization" (Evaluation Gap Working Group, 2006, 73). The authors also referred to the DOE's What Works Clearinghouse's efforts to review the quality of research studies in education, where well-designed RCTs provided the strongest evidence of causal validity (Evaluation Gap Working Group, 2006, 77; cf. chapter three). The report ended with the following statement: "Random-assignment approaches have been demonstrated to be a feasible and rigorous approach to impact evaluation in many situations and should therefore be encouraged and promoted where appropriate" (Evaluation Gap Working Group, 2006, 80).

The group anticipated some common critiques about impact evaluations: "Critics sometimes claim that impact evaluations can only tell whether something has an impact, not why and how" (Evaluation Gap Working Group, 2006, 25). The group countered that a well-done impact evaluation would go beyond the question of whether an intervention had an impact: it would also find out about why and how that impact had occurred. Evaluators would need to have "sound theories and models" and could obtain evidence about the program's mechanisms by collecting data on processes and intermediate outcomes. The group saw an impact evaluation as more than just estimating the average

size of an effect—which is most often the central finding of an RCT. A broader interest seemed to exist assessing "which interventions work under given conditions, what difference they make, and at what cost" (Evaluation Gap Working Group, 2006, 2). They recommended replication of impact evaluations of similar programs in different places, as the most systematic way of increasing the evidence base (ibid., 14). The group argued that well-done impact evaluations would also provide sufficient information about the context to help decide whether findings could be generalized to other situations (Evaluation Gap Working Group, 2006, 14). They did not explain, however, how these generalizations to other situations could be accomplished.

The group pointed out that findings of impact evaluations could be conveyed easily to policy makers (Evaluation Gap Working Group, 2006, 25). Progresa served as an example that policy makers valued the RCT approach and decided to continue funding based on positive RCT findings. The group quoted Julio Frenck, Mexico's former Minister of Health, on the importance of knowledge gained through impact evaluation studies, which could serve as public good (Evaluation Gap Working Group, 2006, 26). They even pointed out that Mexico had passed legislation requiring impact evaluations for a wide range of social development programs (Evaluation Gap Working Group, 2006, 31). The authors also concluded that Progresa's success influenced the design of similar programs throughout the world (Evaluation Gap Working Group, 2006, 18). Ironically, the impact of RCT use on policymaking and program expansion to other countries was not itself tested via an RCT, something the working group did not point out. They made causal statements without the existence of RCT data.

The quest for a global institution to promote rigorous evaluation evidence

The working group painted two scenarios for the future development community, one pessimistic and one optimistic:

> Imagining 10 years into the future… the international community could be in one of the two situations. We could be as we are today, bemoaning the lack of knowledge about what really works and groping for new ideas and approaches to tackle the critical challenges of strengthening health systems, improving learning outcomes, and combating the scourge of extreme poverty.
> Or we could be far better able to productively use the resources for development, based on an expanded base of evidence about the effectiveness of social development strategies." (42-43)

The pessimistic scenario described the status quo of development evaluation, which means groping in the dark. To reach the optimistic scenario, the group called for collective action. To increase the number of rigorous impact evaluations, the working group recommended a voluntary pooled impact evaluation fund (Evaluation Gap Working Group, 2006, 8). Member countries and organizations would contribute funding. Although the cost of an impact evaluation might be too high for an individual organization, a group effort that would distribute costs could fund impact evaluations. This approach would thereby generate valuable evidence and produce knowledge comparable to a "public good." In the long run, ignorance would be more expensive than impact evaluations (Evaluation Gap Working Group, 2006, 23). The working group argued that a pioneering effort by only a few at the vanguard would kindle a larger movement (Evaluation Gap Working Group, 2006, 9; 43). This is reminiscent of Everett Rogers' idea of diffusion, where a few pioneers may trigger an innovation that would later be mainstreamed (Rogers, 1962).

The fund would pursue several goals, including but not limited to: (1) establishing quality standards for rigorous impact evaluations, (2) administering a review process for evaluation designs and studies, and (3) organizing and disseminating information. What the quality standards would look like was left open, but some critics felt that the proposal favored RCTs as the highest standard. Just the idea of a small group of select people putting standards into place was ill-received by some in the development community.

Two working group members from the World Bank, Francois Bourguignon and Paul Gertler, added reservations to the report (Evaluation Gap Working Group, 2006, 44–45). They dissented on issues of how to increase the number of rigorous impact evaluations, and they were especially wary of the nature of the institutional arrangement. Rather than suggesting the addition of another institution, which might increase transaction costs and bureaucracy, they believed in strengthening partnerships between developing country governments, multilateral institutions, NGOs, and researchers. They emphasized the need for capacity building within aid-receiving countries to be able to locally plan and conduct impact evaluations. Ultimately, additional quality impact evaluations would only pay off

if recipient countries would use their findings. The working group co-chairs did not find Bourguignon and Gertler's reservations justified (Evaluation Gap Working Group, 2006, 45).

One year later, the Center for Global Development convened an expert group in 2007 to discuss the charter document for an entity that could fulfill those tasks. The CGD founded the International Initiative for Impact Evaluation (3IE) in 2008, with considerable start-up funds from the Gates Foundation. 3IE started to pool funding from developed and developing countries to pay for impact evaluations since 2009.

On the one hand, people recognized the dearth of high-quality impact evaluations and the need for generating knowledge about what works in development. On the other hand, the publication of the Evaluation Gap report startled the development community, as social scientist Alexandra Caspari observed (Caspari, 2008). They were afraid that the new institution would establish an RCT hierarchy, channeling already sparse funding to meaningless experiments.

In general, the CGD report became a major factor in the RCT debate in international development. It utilized the best development scientists to produce a report that questioned past evaluation practices and recommended a renewed effort in producing rigorous evidence. Although the taskforce members disagreed over the exact methodological approaches, there was a general preference for RCTs in impact evaluation.

3. The Network of Networks on Impact Evaluation: Searching for a compromise on impact evaluation methodology

As part of my research, I investigated how The Network of Network for Impact Evaluation (NONIE) dealt with RCTs by analyzing NONIE's working papers and drafts and attending the NONIE meeting in Washington, DC, in 2008. As a multilateral network comprising communities for whom impact evaluation is imminently relevant, NONIE is a

perfect example of the ambivalence surrounding RCTs. In the following discussion, I present my analysis of the NONIE working papers and drafts, and my observations of the 2008 meeting.

I got involved with NONIE through the American Evaluation Association's Topical Interest Group on International and Cross-Cultural Evaluation (ICCE). During ICCE's annual meeting in November 2007, Zenda Ofir, the former chair of the African Evaluation Association, announced a search for evaluators from developing countries to participate in a NONIE meeting held in Washington D.C. in January 2008. When I stated my interest in impact evaluation, Ofir suggested I contact NONIE's chair, Howard White, who was then a member of the World Bank's Independent Evaluation Group. White suggested I observe the January 2008 meeting—a meeting crucial to finalizing the guidelines on impact evaluation. It would be the third and last meeting of the broader plenum; after which only the steering committee would convene. During the January 2008 meeting, I was able to observe the plenary sessions, subgroup discussions, and hallway conversations. I also collected the distributed documents such as working papers and printed slides. After the January 2008 meeting, I followed up with several NONIE members, including Howard White and Patricia Rogers, the latter of whom became the commentator of the final guidance document. I am indebted to both for their reflections on the NONIE process.

Members of international organizations were compelled to act on CGD's Evaluation Gap report (Caspari, 2008, 138), because they felt uneasy about the proposed new institution prescribing methodological choices for impact evaluation. Instead they wanted to develop their own guidelines based on their prior work in the field. As a result, several multilateral institutions formed the Network of Networks on Impact Evaluation (NONIE) to develop guidelines on high-quality impact evaluations.

Members of NONIE agreed on the need for rigorous impact evaluations. However, they were unable to agree on how to actually best evaluate impact. Should evaluators follow a quantitative decision tree of methods—starting with RCTs and quasi-experiments—or should they rather emphasize participatory evaluations with observational evidence? The draft documents of the NONIE working groups demonstrate the struggle for explicating methodologies that all NONIE members could support. The role of the RCT in the methodology toolbox remained unclear.

More generally, the NONIE example illustrates how the quest for more rigor in development evaluation was accompanied by an ambivalent stance towards RCTs. Members of the international development community felt excluded from the decision-making process and could not support a move toward more quantitative impact evaluations. The question of political dominance partially replaced the original methodological focus.

NONIE background: Responding to the Evaluation Gap report

In November 2006, four months after the Evaluation Gap report's launch, the Organization of Economic Cooperation and Development's Development Assistance Committee (OECD DAC) convened a working group on impact evaluation for its annual meeting in Paris. They formed the Network of Networks on Impact Evaluation (NONIE), which was originally planned as a donor network consisting of the following bilateral and multilateral development institutions: the OECD DAC Evaluation Network, the United Nations Evaluation Group (UNEG), and the Evaluation Cooperation Group (ECG) of the multilateral financial institutions. The World Bank sponsored the secretariat of NONIE, housed at its headquarters in Washington, DC. NONIE's objective was to improve collaboration in the production of relevant and rigorous impact evaluations, to share experiences in conducting rigorous impact evaluations, and to explore the role of impact evaluations in the overall evaluation efforts. These goals were very much in line with the objective of the Center for Global Development's report, but without adding a new

institution (NONIE, November 15, 2006). Another difference was that in addition to focusing on the need for high internal validity by addressing selection bias—which the Evaluation Gap Report had already made its major focus—NONIE also emphasized the need for policy relevance, which included the applicability of findings to other contexts—i.e., external validity.

In a room document from the Paris meeting, NONIE reported that in order to balance pragmatics and rigor, "impact evaluations need[ed] to be well-contextualized and policy relevant" (NONIE, November 15, 2006). Again, NONIE attempted to strike a balance between internal and external validity. They also tried "to maintain a flexible approach, exploiting possibilities for mixed methods whilst retaining technical rigor." Early on, NONIE also emphasized the need to include developing countries as partners and to provide capacity development. NONIE thus took the position of Bourguignon and Gertler, who had dissented in the CGD report suggesting the creation of a new institution for impact evaluation; they called instead for a joint effort between developed and developing partners.

As a NONIE member said during a later meeting: "The Initiative was supposed to be an answer to the Center for Global Development. The three networks [of which NONIE was composed] said that we have done impact evaluation, we bring our expertise together, and we come up with guidelines" (NONIE notes, January 14, 2008). NONIE's plan was to create general guidelines for conducting impact evaluations. To develop these guidelines, the NONIE members formed several subgroups.

NONIE convened a two-day meeting at the World Bank in Washington, DC, in January 2008 to discuss the drafts of the NONIE guidelines developed by the subgroups. I attended this meeting and observed the discussion of the draft documents. The following discussion reports my observations and analysis of this meeting.

In order to follow a participatory evaluation approach and to include evaluators from developing countries in the discussion, NONIE invited twelve representatives from developing countries to their meeting (NONIE notes, January 14, 2008). The twelve

invitees were members of the African Evaluation Association (AfrEA) and the International Organization for Cooperation in Evaluation (IOCE)—i.e., groups that were already active in the evaluation community.

The NONIE meeting preceded The World Bank's conference "Making Smart Policy: Using Impact Evaluation for Policy Making." The conference was a sign that impact evaluation was at the forefront of international development thinking. The fact that one of the invited guest speakers was the CGD's Nancy Birdsall, who had been a lead author of the Evaluation Gap report, showed the importance of the CGD report in the new interest in impact evaluation.[23]

The NONIE "Experimental Group" and the NONIE "Alternative Group"

The NONIE working groups had prepared documents to be discussed during the January 2008 NONIE meeting. The two subgroups working on methodological questions were subgroup 1, which focused on experimental and quasi-experimental designs (henceforth "The Experimental Group"), and subgroup 2, which focused on alternative designs (henceforth "The Alternative Group"). The goal of the meeting was to discuss the draft documents and pull them together into a comprehensive guidance document on impact evaluation. The groups would then present a summary statement during the World Bank's impact evaluation conference.

The Experimental Group authored the documents "NONIE: Impact Evaluation Guidance, Section 1: Introduction" (cited as NONIE Intro, 2008), and "NONIE Subgroup 1: Impact Evaluation Guidance, Section 2: Experimental and quasi-experimental approaches to impact evaluation" (cited as NONIE SG 1, 2008). The Alternative Group authored "NONIE Subgroup 2: NONIE Impact Evaluation Guidance" (cited as NONIE SG 1, 2008).[24] Other documents referred to in this discussion included PowerPoint presentations at the NONIE meetings from the Experimental Group and Alternative Group (e.g., cited as NONIE SG 2 presentation, January 14, 2008), my personal notes

[23] Nancy Birdsall did not attend the conference due to sickness.

[24] These documents were originally uploaded to the NONIE members' page of the World Bank's website www.worldbank.org/ieg/nonie/members.html, but have been removed.

(cited as NONIE notes, January 14, 2008), and the NONIE draft statement distributed at the World Bank conference (cited as NONIE statement, 2008).

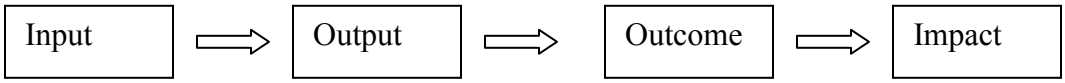
Different emphasis in impact evaluation

The Experimental Group and the Alternative Group both started off with the same definition of impact, provided by the OECD Development Assistant Committee's glossary of evaluation and results-based management (OECD Development Assistance Committee, 2002):

> "Positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended" (NONIE Intro, 2008, 1; NONIE SG 2, 2008, 2)

This definition stands in line with the Logical Framework Approach, where impact symbolizes the last stage of the results chain. The approach is illustrated by the Experimental Group in their introductory draft (FIGURE 3).

FIGURE 3: Results chain from input to impact



Source: NONIE Intro, 2008, 1

An impact evaluation should test the link in the causal chain from input to impact. The focus on impact moves away from the output model (e.g., school attendance) toward an outcome or impact model (e.g., quality of education and quality of life). Outputs are typically the numerical units that inputs created (e.g., numbers of children that attend school). Outcomes are typically defined as short- and medium-term results (e.g., smarter children), whereas impacts are long-term effects (e.g., higher wages as adults). Impact evaluations, therefore, focus on the more long-term effects.

The Experimental Group and the Alternative Group emphasized different aspects of the definition of impact and thus came to different conclusions for choosing methodologies. The Experimental Group focused on the expression "effects produced by." An impact

evaluation would need to tackle the problem of attribution of these effects: How much of the observed effects can be attributed to the intervention itself (NONIE SG 1, 2008, 1)? The Experimental Group thus framed impact as "net effect," similar to the Evaluation Gap Working Group's definition of "net impact."

The Alternative Group focused on the multi-faceted, non-quantitative nature of impact, including the full range of impacts at all levels of the results chain. They quoted the second part of the OECD DAC definition:

> "These effects can be economic, socio-cultural, institutional, environmental, technological or of other types (NONIE SG 2, 2008, 2).

The Alternative Group specifically referred to the levels of families, households, and communities, as well as institutional, technical, and social systems. They also emphasized the long-term, complex nature of impact. As a result they found the RCT—a simplistic methods in their view—an inappropriate tool for most impact evaluations.

Although both groups started with the same definition of impact, agreements on the concept of rigorous impact evaluations were hard to reach. One key question was to what degree an impact evaluation actually relied on a counterfactual, which could be represented by, for example, a randomized control group or comparison group.

The Experimental Group's counterfactual response to impact evaluation

The Experimental Group argued that that any evaluation of impact required a counterfactual statement of effects produced by the intervention—i.e., effects that would not have been observed in the absence of the intervention. Therefore, how to address the counterfactual would be the central issue in impact evaluation design (NONIE Intro, 2008, 1). In another section, they wrote that: "The NONIE guidance is concerned with impact approaches which establish a counterfactual" (NONIE Intro, 2008, 2).

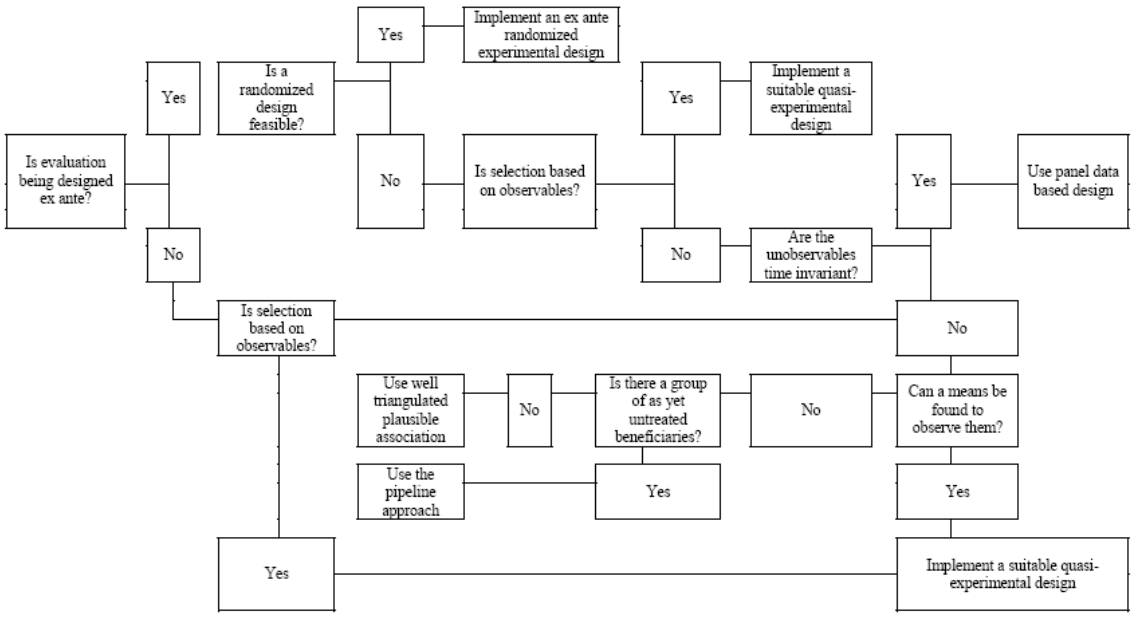
The Experimental Group conceived of the counterfactual as how the situation would have developed without the intervention (NONIE SG 1, 2008, 1). By comparing the factual and counterfactual situation, they would be able to attribute the net effect to the intervention. Such a comparison between factual and counterfactual was challenging,

because, as the subgroup reported: “It is not possible to observe how the situation would have been without the intervention” (NONIE SG 1, 2008, 1).

For the Experimental Group, the most common solution to isolate the net effect was the selection of a comparison group (NONIE SG 1, 2008, 2). The evaluators needed to be careful that the comparison group would exhibit characteristics similar to the treatment group, minus the influence of the intervention (NONIE SG 1, 2008, 3–4). The Experimental Group argued that the statistical chance for biased assessments would be small when the treatment and the control group were selected randomly (NONIE SG 1, 2008, 4). At the same time, they acknowledged that such “true experimental designs” had been rare in development settings. They referred to medical tests that routinely use the RCT approach (NONIE SG 1, 2008, 4).

The Experimental Group suggested using a decision tree to select the appropriate evaluation approach for addressing selection bias (NONIE SG 1, 2008, 6; cf., FIGURE 4).

FIGURE 4: Decision tree for selecting an impact evaluation approach



Source: NONIE SG 1, 2008

If an evaluation were to be designed before an intervention had started, the first question would be whether an RCT would be feasible. If not, they suggested other designs. In the event the selection of treatment group is based on observable differences, they recommended quasi-experimental designs. If unobservable differences are time invariant, they recommended panel data-based designs. If the differences between the treatment and control group are unobservable and could not be estimated, they found that the problem of selection bias could not be addressed (NONIE SG 1, 2008, 6). Instead, they suggested "plausible association" via program theory and triangulation. However, the Experimental Group would not consider this design a rigorous impact design. Such a design would fall under interpretive-qualitative approaches.

The Experimental Group envisioned a decision-tree where the RCT would be the top choice if feasible, whereas qualitative approaches would be the bottom choice if none of the quantitative techniques worked. Still, the Experimental Group maintained that so-called rigorous impact evaluations would not be restricted to experimental and quasi-experimental approaches. (Quasi-)experimental approaches would only apply for interventions with many observations. However, when the number of observations was small—such as in cases of institutional or national policy change—quantitative and statistically-based approaches would no longer be meaningful. Instead, they suggested a modeling-based approach and a narrative approach along the lines of historical research (NONIE Intro, 2008, 5). The Experimental Group did not explicate these non-quantitative approaches. Had they provided further elaboration on these methodologies, they might have more successfully demonstrated their commitment to including alternative non-quantitative methods.

An impact evaluation would need to answer not only the question of whether the program works, but also the question of why there was or was not an impact (NONIE Intro, 2008, 3). Answering the why question would help generalize from a specific context, and it would also help in making decisions about scale-ups and how to improve interventions (NONIE SG 1, 2008, 5). In fact, impact evaluations that focused merely on measuring impacts would be mere "black box evaluations," without any indication as to why the intervention had the expected impact (see also Bourguignon & Sundberg, 2007). In order

to open the black box, the Experimental Group suggested mixed-methods evaluations that included qualitative and quantitative information (NONIE SG 1, 2008, 6).

For the Experimental Group, evaluators would need to establish a program theory in order to answer the "why question." Such a theory-based evaluation design would reconstruct the logic behind a program and identify the channels through which the program was expected to operate (NONIE SG 1, 2008, 5). Program theory would explicitly capture how, why, and for whom an intervention should generate a certain impact (see also Caspari, 2009, 206). To reconstruct a program theory, evaluators would rely on the review of planning documents, such as Logical Framework Approaches. Rigorous impact evaluations would finally be able to test the assumptions of Logical Framework Approaches, which had been used since the early 1970s.

A rigorous impact evaluation combined with a theory-based analysis would increase the relevance of an evaluation for policy makers. This echoed the opinion of Howard White, a member of the Experimental Group, that "there is not necessarily a tradeoff between what is called here rigorous impact evaluation—i.e. one which applies the appropriate technical procedures—and relevance, which depends on a well-contextualized theory-based approach" (White & Barbu, 2006, 2).

Overall, the Experimental Group found that any impact evaluation would need to explicitly deal with methodological problems when measuring impact (NONIE Intro, 2008, 2). Especially at issue was the selection bias. Evaluators could use a variety of methods, albeit preferably quantitative, to test the validity of the program's theory-based assumptions and to analyze the various links in the intervention chain (NONIE SG 1, 2008, 5). The Experimental Group stressed the importance of mixed-methods evaluations that included both qualitative and quantitative information (NONIE SG 1, 2008, 6).

In sum, the Experimental Group was not interested in the polarization of quantitative approaches versus qualitative approaches. However, they found that qualitative approaches alone would not be able to address the problem of selection bias. They saw a

need for mixed-methods evaluations to explain why and how interventions worked, and left it to the Alternative Group to illustrate these qualitative evaluation approaches.

The Experimental Group's arguments were more diversified than the Evaluation Gap Working Group's report regarding methodological choices. They provided a decision tree on how to select evaluation designs, which was based on the timing and the available data for an intervention. The decision tree, however, was only limited to interventions that contained many observations (i.e., units under study), which is often not the case for institutional policies or country-wide interventions. Finally, although the decision tree was not intended as a hierarchy of methods, it was still close to hierarchical thinking. This approach was something that the Alternative Group did not appreciate this, given their preference for methodological pluralism.

The Alternative Group's factual response to impact evaluation

In the foreword, the Alternative Group's report on alternative designs made the following statement:

> "SG 1 has provided an introductory document [i.e., NONIE Intro, 2008]. SG2 submits that this has a narrow perspective and does not address the comprehensive view of impact evaluation that is espoused by NONIE members. SG2 thus presents an alternative proposal for Section 1 of the NONIE Guidance Document (See Section 1 of the SG2 document). This presents a multi-faceted and contextual character of impact evaluation in development contexts." (NONIE SG 2, 2008, i)

The Alternative Group set a confrontational tone with these words. They accused the Experimental Group of being too narrow in their understanding of impact evaluation and suggested a more comprehensive view of impact evaluation. For one, they did not want to restrict themselves to methodological issues. They also wanted to address issues on ethics, implementation, and management of impact evaluations. Second, they found the Experimental Group's definition of impact to be too narrow. Recall that both subgroups started with the same definition of impact by the OECD DAC glossary (cf., above). The Experimental Group emphasized the idea of the counterfactual and net effect. The Alternative Group emphasized the complex nature of impact. They did not find the concept of a counterfactual a necessary part of impact. They also did not subscribe to the concept of attribution—i.e., whether changes could be attributed to the intervention—

because they felt attribution generally implied a counterfactual. Instead they preferred the concept of contribution. They argued against the use of an "explicit counterfactual," which they equated with the use of randomized control groups or RCTs. Instead, the Alternative Group used the concept of a causal contribution analysis, which would rely on qualitative approaches.

The Alternative Group observed that many development interventions were non-standardized, emergent, and complex. Such interventions included community development, natural resources management, and emergency situations (NONIE SG 2 presentation, January 14, 2008). Emergent outcomes would make pre-identification of outcomes difficult (NONIE SG 2, 2008 23). Therefore, complex interventions would often not allow for an explicit counterfactual via a randomized control group. For evaluation purposes, no counterfactual would be necessary. The Alternative Group quoted the historical sciences and the natural sciences, which generated evidence without the use of a formal counterfactual (NONIE SG 2, 2008, 16).

A complex intervention would need a multi-disciplinary, mixed-methods evaluation approach that established rigor via triangulation (NONIE SG 2, 2008, 3; 14). The Alternative Group rejected any idea of a decision tree. They found that "methodological appropriateness should be considered the 'gold standard' for impact evaluation" (NONIE SG 2, 2008, ii), a statement that the Experimental Group could probably also accept.

Each impact evaluation would test a theory against "logic and the evidence available on the various assumptions behind the theory of change." Evaluators would identify possible explanations for the program impact and then rule these out, based on observational evidence (NONIE SG 2, 2008, 24). Sources for plausible program theory could be expert opinions, beneficiaries, and research (NONIE SG 2 presentation, January 14, 2008).

The Alternative Group deplored that the Experimental Group provided what they felt was an overly simplistic response to the need for policy evidence, because experimental and quasi-experimental approaches would only capture "what works, based on numbers and statistics" (NONIE SG 2, 2008, 4). Impact evaluations, however, would need to use more

complex models to address impact from various angles. The Alternative Group relied on the concept of "realistic evaluation," which Pawson and Tilley had developed in the 1990s in the United Kingdom. Realistic evaluation asked more comprehensive questions, including: "What works for whom in what circumstances and in what respects, and how?" (NONIE SG 2, 2008, 25). A rigorous impact evaluation would address all these questions. Only then could it guarantee both internal validity and external validity, i.e., generalizability and transferability to other policy contexts (NONIE SG 2 presentation, January 14, 2008).

The Alternative Group's arguments were important because they opposed a narrow concept of impact evaluation as measuring merely net effect. At the same time, the Alternative Group did not adequately value the concept of the counterfactual and attribution. In line with what one NONIE member emphasized, I argue that the counterfactual is an imaginary notion, and that in fact, every impact evaluation would have to deal with such an imaginary or implicit counterfactual (NONIE notes, January 14, 2008). Moreover, there was no substantive distinction between attribution and contribution, the main difference being in the emphasis of other simultaneous effects when using the term "contribution." But even contribution analysis would deal with cause-effect questions when evaluating impact. The Alternative Group also did not explicate alternative methods in detail to assess and understand program impact.

The 2008 NONIE draft statement and its controversy

NONIE encompassed a variety of opinions and conceptions of what impact evaluation was. The Experimental Group emphasized mostly methodological questions about selection bias. The Alternative Group was concerned with qualitative approaches and understanding impact evaluation in a larger development context. Finding a unifying voice was a difficult undertaking. The final plenary session of the NONIE meeting in January 2008 illustrated the heterogeneity of opinions and inability to reach a satisfactory consensus.

Before the final session, the NONIE leadership had circulated a two-page summary, "Draft NONIE Statement on Impact Evaluation" (NONIE statement, 2008). The first

paragraph was concerned with the need for quality impact evaluations, especially in the context of the Millennium Development Goals and results agenda. In the second paragraph, the summary quoted the OECD DAC definition of impact. Accordingly, quality impact evaluations would focus on long-term effects, including negative, indirect, unintended and secondary effects. The evaluations would thus encompass more than just positive, direct, intended and primary effects—which would be more easily measured. However, the summary did not include the second part of the definition of impact, which concerned the economic, socio-cultural, institutional, environmental, and technological effects. It was this latter half that had been of greater concern to the Alternative Group.

The draft stated that impact evaluations would be issues-driven and priority-driven. Therefore, they would use appropriate methods. This meant that they would not choose interventions based on their amenability to certain impact evaluation approaches. This was an indirect reference to choosing only interventions to be evaluated that would allow for RCT methodology. The consultation draft of the Center for Global Development had stated that half of the impact evaluations had to be RCTs—which would not have tailored the method to the topic under study.

The NONIE draft statement then listed the available methodological choices in the following order: randomized controlled trials, regression discontinuity, propensity score matching, and other regression-based approaches that deal with problems of sample selection. The statement further acknowledged that these quantitative approaches could not be used widely for the type of existing interventions. Approaches would therefore need to include other methodologies to cover the full range of development interventions including sector-level and country-level interventions. Unfortunately, the NONIE statement did not include what those methodologies should be.

The summary draft did not emphasize any hierarchical order of these quantitative approaches, as the Experimental Group had implied in their decision tree. The statement did not include qualitative methodologies as stand alone for impact evaluation. The NONIE statement, however, specifically promoted the use of mixed methods. The statement referred to “quantitative and qualitative data,” thus being mostly concerned

with how the data were collected rather than how they were analyzed. The statement did not make a clear distinction between data collection and data analysis methods.

On the one hand, the draft statement emphasized the problem of biased sample selection that would bias impact estimates, which would decrease internal validity. On the other hand, NONIE also advocated a "theory-based evaluation design" according to Pawson and Tilley's principles of realistic evaluation (Pawson & Tilley, 1997). This approach was much more concerned with external validity: "under which circumstances findings of the evaluation are transferable to other contexts." (NONIE Statement, 2008) A theory-based design would establish a hypothesis on the process of an intervention from its inputs to outputs, outcomes, and impact (cf., the Logistical Framework Approach in Figure 2, and the results chain in FIGURE 3). Thus, an impact evaluation would move beyond the question of merely *whether* an intervention had its intended effect—which had been a major focus of clinical RCTs in drug development. Theory-based designs would also address the "question of why—or why not—an intervention had the intended impact." Opening the black box of development of particular interventions would help clarify the mechanism of an intervention via its result chain and thus could offer clues whether it could be replicated in other contexts. Thus, an impact evaluation would contribute to policy relevance, because policy makers were interested in how and why interventions worked.

The NONIE summary also emphasized the collaborative, participatory partnership of evaluators from developing and developed countries. NONIE supported a "Southern-led evaluation program" (NONIE statement, January 5, 2008). NONIE's decision to include members from developing countries in the development of impact guidelines showed their commitment in Southern-Northern partnerships.

When the NONIE leadership presented the draft statement, several NONIE members expressed their discontent (NONIE statement, January 5, 2008). One member argued: "The statement refers to a wide variety of potential methods, but it is implied that there is a hierarchy of methods. It does not quite say that, but it is implicit" (NONIE notes, January 14, 2008). The person requested the removal of all the quantitative methods from

the list, because their inclusion could make a false impression regarding NONIE's mixed-methods stance. He also suggested making the "mixed-methods" stance more visible in the statement.

Referring to the subgroup documents, one NONIE member poignantly raised the question as to how NONIE could rapidly come to a principal agreement, as it would be naïve to assume that they had a finalized document in hand (NONIE notes, January 14, 2008). Another person from a developing country demanded a new statement to address the various gaps identified by the Alternative Group and which would go beyond the methodological issues. She felt the statement to be too one-sided and too premature to be released in its current form. Members in general felt that the NONIE statement was preliminary. NONIE itself was regarded as a lightly structured, fledgling organization, still forming in the next decade.

The NONIE meeting ended with the agreement of hiring an external consultant to create a compromise between the subgroup documents. The dissatisfactory conclusion of NONIE's January 2008 meeting did not change for the better in the following year. In fact, divisions grew and a compromise proved to be unreachable. The inclusion of a wider group of evaluators, including representatives of national evaluation associations and representatives from developing countries, revealed that impact evaluation was not a clearly defined topic, and the role of the RCT in impact evaluation remained unclear.

The 2009 NONIE guidance document: A contested compromise

The NONIE steering committee asked Frans Leeuw, who was from the Maastricht University in the Netherlands, to produce the final draft of the NONIE guidelines based on the subgroups' individual documents. The committee also invited Patricia Rogers from RMIT University in Australia, who had advised NONIE's Alternative Group, to comment on Leeuw's draft.

Leeuw and his collaborator Jos Vaessen presented their draft at the international Impact Evaluation Conference in Cairo, Egypt, in April 2009. Up front, the guidance document added the following caveat: "The document should not be taken to represent the agreed

positions of all the individual NONIE members. The network membership and the authors recognize that there is scope to develop the arguments further in several key areas."[25] Because member perspectives on the definition, scope, and appropriate methods of impact evaluation differed widely, the final NONIE guidance merely represented the views of the authors commissioned by NONIE, Leeuw and Vaessen.

The nine-page executive summary, drafted by Arup Banerji from the World Bank, outlined the methodological approaches to impact evaluation (Leeuw & Vaessen, 2009, ix). Banerji referred first to the agreed-on definition of impact, in accordance with the OECD DAC glossary. Then he focused on two concepts of impact evaluation: attribution and counterfactual—both of which were in the original NONIE Experimental Group draft (NONIE SG 1, 2008). In line with the Alternative Group, the guidance stated that: "No single method is best for addressing the variety of questions and aspects that might be part of impact evaluations" (Leeuw & Vaessen, 2009, x). Nor, they argued, was there a "gold standard, in the sense of a single method that is best in all cases" (Leeuw & Vaessen, 2009, xiii). However, depending on the scope, objectives, and design of the intervention, some methods would have a "comparative advantage over others in analyzing a particular question or objective" (ibid., x, xiii). For "single-strand" interventions, experimental and quasi-experimental evaluation designs might have a comparative advantage in assessing attribution of causal effects (ibid., x, xiii)—a statement which the Experimental Group would have supported. On the other hand, programs deemed "complicated" and "complex" with extensive range and scope would not be suited for narrow "counterfactual estimation" by RCTs or quasi-experiments (Leeuw & Vaessen, 2009, xiv). Regression-based approaches could be used instead. The authors characterized non-quantitative techniques as being less effective in many of the cases addressing attribution—an opinion that the Alternative Group would have disputed.

Non-quantitative techniques would help in the theory-based unpacking of institutional and beneficiary levels of impact. A non-quantitative, theory-based approach would help stakeholders understand an intervention. The evaluators would first need to develop a theory of intended or implicit objectives, and then identify the sometimes tacit social,

25 http://www.worldbank.org/ieg/nonie/guidance.html, retrieved September 7, 2011.

behavioral, and institutional assumptions these objectives entailed (Leeuw & Vaessen, 2009, xii). The evaluation would then test those assumptions by either reconstructing the causal "story" or by formally testing those assumptions. Qualitative methods would help with "construct validity" by ensuring that the variables being measured adequately represented the underlying realities of development interventions (Leeuw & Vaessen, 2009, xv). A mixed-method triangulation of data collection and analysis would increase internal, external, and construct validity (Leeuw & Vaessen, 2009, xvi). Overall, the authors concluded that, "well-designed quantitative methods are usually preferable for addressing attribution and should be pursued when possible" (Leeuw & Vaessen, 2009, xv). For single-strand, relatively homogeneous interventions, the authors described RCTs as "better than most other methods in terms of *internal* validity," based on a clearly identified counterfactual (ibid., xv). *External* validity would rely on systematic repetition of RCTs across a range of settings and policy options.

Overall, both the executive summary and the main document attempted a compromise between the two NONIE subgroups, but ultimately they were unable to fully integrate the Alternative Group's call for non-quantitative methods to establish causal effects. Like the Experimental Group, Leeuw and Vaessen called for establishing a counterfactual. They used the example of installing water pumps as a rare case, where a factual evaluation would suffice to capture impact (Leeuw & Vaessen, 2009, 21). Normally, controlling for other influences would require a control group, which would simulate such a counterfactual. Random assignment to the participant and control group would be considered the best way to create equivalent groups (Leeuw & Vaessen, 2009, 22). The second best alternative would be matching techniques by creating control groups similar to the participant groups (ibid., 23). Labels such as "best" or "second best" were reminiscent of the decision tree used by the Experimental Group. From a counterfactual perspective, the RCT was the best evaluation design, but not necessarily the most feasible. The authors pointed out that counterfactual evaluations could not be implemented in full-coverage interventions, such as price or environmental policies. Qualitative methods, however, would have the drawback of not quantifying effects attributable to an intervention (ibid., 31). They often could provide a framework in which quantitative methods could be used.

In a personal communication, Patricia Rogers stated that the authors did not integrate her concerns about RCTs or her comments about alternative designs into the final document (Rogers, personal communication, April 29, 2010). Rogers was disappointed about the process of finalizing the NONIE guidance and felt that her voice and the voice of her colleagues from the Alternative Group had not been adequately reflected.

During a NONIE steering committee meeting, Andrew Warner from the NONIE secretariat pointed out that NONIE's evaluators came from "different intellectual subcultures" (Warner, presentation, October 3, 2008). He noted that some disputes were "really more about words than anything else," while others were rooted in misunderstandings or partial understandings about other methodologies. There was, however, a "remaining core of real disputes on substantive issues" (Warner, presentation, October 3, 2008), which I will further explore in the final chapter where I compare arguments of RCT supporters and RCT critics. I found that semantics differed in how people used central terms such as "impact evaluation," "counterfactual," and "mixed methods." I argue that even the concept of an RCT suffered from partial misunderstandings, especially surrounding the word "controlled." Did "controlled" mean just purely statistical control, or was some other control exercised in an RCT? For several evaluators, the term "control" implied external control, and thus was suspect from a participatory evaluation tradition.

In sum, the NONIE-internal struggles showed that disagreement still existed concerning the role of RCT designs in development evaluations. On the one hand, one group called for rigorous methods based on RCTs wherever feasible. On the other hand, qualitative-participatory evaluators argued for methodological pluralism to adequately determine program effects and underlying processes. They were also concerned that the exclusive focus on methodological questions would overlook issues of policy relevance and political inclusion.

4. European Evaluation Society's criticism of RCT primacy

The Center for Global Development's report also triggered a response from the European Evaluation Society (EES). In December 2007, EES released a statement regarding "the importance of a methodologically diverse approach to impact evaluation – specifically with respect to development aid and development interventions." And in April 2009, EES released "Comments on the Draft NONIE Guidance on Impact Evaluation." In both cases, EES shared disappointment about the increasing preference given to randomized controlled trials (RCTs) for evaluating impact.

In the 2007 statement, EES denounced one contemporary perspective being strongly advocated, which was that "the best or only rigorous and scientific way of doing [impact evaluations] is through randomized controlled trials (RCTs)." Instead, EES supported "multi-method approaches" and would not consider "any single method such as RCTs as first choice or as the 'gold standard'" (ibid., 1).

What, specifically, did EES mean by "method" and "RCT"? The very expression "method," derived from the Greek word for 'way' (met-odos), is an ambiguous term with various meanings depending on context, though it most frequently refers to ways of data collection, data analysis, and overall function. First, EES seemed to regard the "RCT" as an exclusive method—one either uses it or one does not. I would argue, in contrast, that the RCT is a sample-generating method, which should be combined with multiple methodological tools for data collection, analysis, and interpretation. Secondly, the EES statement seems to equate the RCT with a laboratory experiment, where the experimenter and the environment exert control over the subjects and are hence "controlled" trials (cf., Table 9, point 2). This is a misleading view of RCTs. The RCT does not exercise control over the study subjects by controlling their environment; it controls via statistical control, which is anchored by the random assignment process. The RCT would therefore allow for less environmental control because it statistically controls exogenous factors (cf., chapter 6). This was a fundamental insight by Ronald Fisher.

EES criticized the RCT along several lines. The following eight points summarize the EES claims (cf., Table 9).

TABLE 9: EES's arguments against RCTs to determine development impact

An RCT …	(1) is not able to assess complex, nonlinear interventions with multiple causes;
	(2) needs to rigorously control for context and other intervening factors;
	(3) is unable to adapt interventions during the evaluation process;
	(4) is unethical at times;
	(5) lacks generalizability/external validity and thus prohibits representativeness for scaling-up;
	(6) promotes black-box evaluations without understanding the intervention process (i.e., it does not understand "what works for whom and under what circumstances");
	(7) does not focus on unintended, unanticipated outcomes; and
	(8) does not focus on participatory evaluation.

Compiled and adapted by the author

Table 9 highlights the EES criticism that RCTs would not allow for adapting interventions during the evaluation process. In reality, program administrators often had to adapt development interventions to the changing political or economic environment, such as civil war. Medical trial literature, however, has questioned this inability of RCTs to adapt. Clinical trials have been using adaptive models (Cook, 2007). A further critique was that RCTs would focus only on the desired short-term effects, but they might not capture more comprehensive, long-term, unintended, and unanticipated effects. RCT baseline data may not capture unintended effects, and thus would not be used to determine the program's net effect. The issue of how to define effect had played an important role in NONIE's process of establishing guidelines for rigorous impact evaluations. In general, several arguments brought forward by EES were valuable in as much as they pointed out RCT limitations—limitations that could be addressed via adding other methodological approaches.

As point (8) indicates, the statement follows the participatory evaluation tradition such as the Participatory Rural Appraisal (PRA) developed by Robert Chambers in the 1980s, combining various activist participation and observation tools (Chambers, 1994). Participatory approaches proposed that any intervention impact was temporally and culturally situated (Mayoux & Chambers, 2005, 35). Thus, local beneficiaries had critical

knowledge of program implementation and effects, and could therefore make valuable contributions to the evaluation process (Mayoux & Chambers, 2005, 25).

The "analytical capabilities of local people" were important in generating reports on program impact (Chambers, 1994, 953). Participatory approaches, such as PRA countered seemingly top-down, standardized, and blueprint models of evaluations, based on unreliable data from questionnaire surveys. Participatory evaluations would empower local people, would be more exploratory in nature, and would be adaptable to the local context. They originally deemphasized formal methods, and instead encouraged personal judgment (Chambers, 1994, 959). Participatory approaches partially arose out of a political and transformative purpose—a purpose specifically intended to make the poor capable, powerful, and self-validating (Mayoux & Chambers, 2005, 29, 36; Chambers, 1994, 963). Participatory evaluation was an important reference point of the EES statement as well as NONIE's Alternative Group. Contrarily, RCT evaluators from Western countries would bring a flair of "colonialism" to the developing world (cf., Duncan, 2008). Furthermore, any attempt to establish a rigorous hierarchy of methods was found suspicious of being exclusionary.

The 2009 "EES Comment on the Draft NONIE Guidance on Impact Evaluation" supported the original two-page NONIE statement from January 2008 and pointed to the Guidance's deviation from the original idea of methodological diversity. Even if the NONIE statement was devoid of any methodological hierarchy—though, recall that one NONIE member pointed out that the methods list started with RCTs—the Experimental Group's document contained a decision tree of what methods to use. The RCT was the first choice. The NONIE Guidance no longer incorporated such a decision tree, but pointed to the comparative advantages of each method. EES seemed to be overcritical of the NONIE Guidance, partially because it had felt under-represented in the feedback loop.

Both EES documents assumed an egalitarianism among methods, and they alerted evaluators to take into account alternative approaches and methods for producing policy-

relevant impact evaluations. They did not, however, further explicate these alternative approaches and methods.

5. USAIDs new evaluation policy—attempting to strike a balance

The new evaluation policy of U.S. Agency for International Development (USAID) in 2011 reflected a move towards both using RCTs and mixed methods. Ruth Levine, who had co-led CGD's Evaluation Gap Working Group, became the director of evaluation at USAID in 2008. Levine's goal was to make USAID the frontrunner development agency, as it had been in the 1970s (Levine, presentation, April 28, 2010). Recall that USAID developed the Logical Framework Approach in 1971, which many bilateral and multilateral agencies proceeded to implement in subsequent years up until the current day.

USAID's new evaluation policy (2011) attempted to promote both experimental designs and qualitative approaches. The Evaluation Policy Task Team took RCT criticism into account and argued for evaluations based on best methods (U.S. Agency for International Development, January 2011):

> **"Based on the best methods:**[26] Evaluations will use methods that generate the highest quality and most credible evidence that corresponds to the questions being asked, taking into consideration time, budget and other practical considerations. Given the nature of development activities, both qualitative and quantitative methods yield valuable findings, and a combination of both often is optimal; observational, quasi-experimental and experimental designs all have their place. No single method will be privileged over others; rather, the selection of method or methods for a particular evaluation should principally consider the empirical strength of study design as well as the feasibility (9)."

The statement demonstrated a well-balanced view of qualitative and quantitative methods in evaluation, respecting the value of each method in evaluation and supporting a combination of those. The statement seemed to reflect EES's quest for methodological pluralism without hierarchical thinking, as no method should be privileged over others. Note, however, that this statement was not limited to impact evaluation, but referred to evaluation in general. The policy's tone changed when it referred to impact evaluations.

In the tradition of CGD and NONIE (Experimental Group), the USAID group argued that any impact evaluation needed a rigorously defined counterfactual. For impact evaluations, they argued that "experimental methods generate the strongest evidence" (9). Put another way, evaluations where "beneficiaries are randomly assigned to either a 'treatment' or a 'control' group provide the strongest evidence of a relationship between the intervention under study and the outcome measured" (4). As a result, when determining impact, "alternative methods should be utilized only when random assignment strategies are infeasible" (4). These statements revealed a clear primacy of the RCT over any other method for impact evaluation.

In sum, the RCT thinking in international development arose from a general disappointment in evaluation quality. The Logical Framework Approach of 1971 addressed the need for impact-oriented thinking in development. Development agencies, however, were not able to measure whether they had truly achieved the desired impacts. The 2006 Center for Global Development's Evaluation Gap report deplored the dearth of rigorous impact evaluations and suggested forming a council to promote them. The multilateral and bilateral institutions responded with the Network of Network on Impact Evaluation (NONIE), which attempted to develop guidelines for quality impact evaluations. While the members agreed on the definition of impact, they could not agree on how this impact would be best measured. One working group suggested a decision tree, which started with the RCT as the best method, while another group wanted methodological pluralism. NONIE did not author the final Guidance document. The European Evaluation Society negatively responded to RCT movements and suggested methodological diversity. The new USAID Evaluation Policy incorporated both traditions: RCT-preference for impact evaluations, and multi-methods approach for evaluations in general.

This discussion of RCTs in international development shows how the methodological pendulum has moved from the RCT-only side toward a more inclusive approach in order for evaluations to become more policy relevant. Conditional cash transfer evaluations, for

---

[26] Emphasis in original text

instance, often include a qualitative-observational study to understand the underlying processes of the intervention, which are important for generalizing findings. Furthering this line of argumentation in the policy recommendations (cf., Chapter 5), I point to the often-unacknowledged fact that RCTs always incorporate many qualitative-interpretive components, which evaluators and policy makers need to understand to make sense of RCT findings and put them into the right perspective.

CHAPTER 5: THE RCT MODEL IN COMPARATIVE POLICY PERSPECTIVE AND ITS CHALLENGES

What lessons for program evaluation can be discerned from the use of RCTs in three distinct areas of policy making? How can these lessons inform public policy evaluation? To answer these questions, I first provided a nuanced analysis of how the RCT movements emerged and developed individually in the policy fields of medicine, education, and international development and how they encountered and dealt with opposition (cf., chapters 2, 3, and 4).

In this chapter, I first compare policy perspectives, and I demonstrate that the seemingly disparate policy fields all privilege the RCT. However, this privileged status is more easily accomplished in medicine than in education or international development due to the nature of the evaluation. Still, even medicine has encountered a countermovement criticizing the sole reliance on RCTs—a countermovement whose advice education and international development might be well-advised to heed.

Second, I analyze the different perspectives among RCT supporters and RCT critics across the three policy fields in order to construct a scaffolding for productive discussions of RCTs. As I show, epistemological disagreements and interpretive differences in terminology –even the term RCT—fueled the debate over the primacy of RCTs. I argue that achieving a consensus on terminological use of concepts would assist in bridging the RCT supporters' and critics' understanding of the RCT model.

Third, I identify challenges of implementing the RCT model across all three policy fields, again referring back to chapters 2, 3, and 4. I argue that the fields of education and international development may not just point to the strengths of the RCT model in medicine, but also highlight the need to closely investigate its challenges—challenges that are even more pressing for education and international development than for medicine. For example, one of the most central factors that RCTs are meant to address—heterogeneity—does not only apply to individual subjects such as patients, but also to households, education systems, and the policy context in general. A policy maker needs

to understand the challenges of using the RCT model, such as how to deal with the problem of insignificant RCT findings. They may further benefit from understanding how challenges associated with RCTs were resolved across policy fields. For example, the medical field reduces the number of insignificant findings by conducting more pre-clinical, pre-RCT research. Understanding these challenges helps evaluators and policy makers to understand the appropriate application of the RCT model by neither overestimating nor underestimating its power and use.

My comparative analysis leads to policy recommendations that foster a sound understanding of the RCT model and what its place is in the knowledge and policy generation process. I argue, for example, that evaluators and policy makers need to understand possible biases in the RCT model (cf., chapter 2), the qualitative-interpretive reasoning surrounding the RCT process, and how representative findings really are. I also recommend broadening the evidence-base beyond the RCT model through, for example, utilizing cost-effectiveness research when evaluating impact. By doing so, RCT evaluations may become more relevant for the particular policy context.

Lastly, I conclude this chapter by offering my view of limitations of the study and directions for future research. Future research would benefit from, among other things, a similar analysis for non-experimental methods and how they could contribute to evaluating impact in public policy.

1. Comparative perspective of RCT movements in medicine, education, and international development

Comparing the RCT movements across the three policy fields of medicine, education, and international development reveals many parallels beneath the idiosyncrasies of each field. Table 10 summarizes a comparison of the RCT movements across the three policy fields. Note that medicine appears to be the frontrunner, with education and international development following. The recent references of these latter fields to the "higher status" medicine and its strong use of the RCT model suggest some modeling after the medical RCT approach.

TABLE 10: Comparison of RCT movements in medicine, education, and international development

Field / ***Topic***	*Medicine*	*Education*	*International Development*
Adoption of RCT approach	1940s U.K. trials (Streptomycin)	1970s federal programs	1960s health and nutrition studies
Renewed interest in RCTs	1992 Evidence-Based Medicine	1998 National Reading Panel 2001 NCLB	2003 J-PAL 2006 CGD report
Institutionalization of RCT approach	1970: 2 RCTs per drug required (FDA)	2005: Competitive preference priority (USDOE)	2011: Preference in impact evaluation (USAID)
Countermovements	Medical researchers and clinicians	Evaluators and school practitioners	Development evaluators and practitioners
Adjustments	Comparative effectiveness research	WWC guidelines on single-case studies	USAID evaluation policy

Adoption of the RCT approach

In the field of medicine, the U.K. Medical Research Council promoted randomized medical trials in the 1940s through studies such as the Streptomycin trial of 1948. Soon thereafter trials were popularized in the United States, culminating in the Poliomyelitis trials in the 1950s. By utilizing public schools, the Polio trial became a "public experiment." The desperate need for a vaccine and the positive results of the trial helped foster the general acceptance of RCTs as an evaluation tool—a tool that would modernize medicine into a lifesaving profession.

By looking up to the medical sciences as the frontrunner in designing and implementing RCTs, scientists in education and international development recognized RCTs as *the* modern methodological approach. RCTs of health and nutrition interventions entered the field of international development in the 1960s (e.g., the Guatemalan Villages study). The U.S. Government also sponsored large-scale education trials in the 1970s and 1980s, such as the Head Start experiments.

It was medicine that paved the way for RCTs to be accepted as scientific methodology and modernized means of knowledge generation. So-called lower-status policy fields

such as education and international development subsequently adopted the RCT methodology to transform themselves into more scientific disciplines. What they did not realize is that large-scale RCTs in medicine are only a small slice of knowledge generation. The drug development process utilizes large-scale RCTs only in the third phase. Education and international development did not put a similar system into place that required and valued pretrial research. They therefore may have introduced RCTs too prematurely as an evaluation tool in the knowledge generation process.

Renewed interest in RCTs

The Evidence-Based Medicine movement in 1992 was an attempt to bring RCT evidence to the medical practitioner's office (Evidence-Based Medicine Work Group, 1992). Also beginning in 1992, the Cochrane Collaboration produced systematic reviews, which used an RCT hierarchy of evidence to evaluate existing research in medicine. They considered the RCT the optimal evaluation design, and any quasi-experimental and non-experimental designs were discounted against the RCT standard.

In the late 1990s, stakeholders in U.S. education and international development had become dissatisfied with the available evidence base of effective programs and found that past evaluations often had justified program interventions without truly establishing the causal attribution of program effects. Modeled after Evidence-Based Medicine, the National Reading Panel established experimental standards of evidence in 2000. Moreover, the federal Institute of Education Sciences established the What Works Clearing House in 2003, which promoted RCTs as the default approach in educational impact evaluations. A main problem in education was that most of the previous research were not RCTs and therefore would not be included in the review process, but discarded up front.

In international development, the renewed RCT movement started with the establishment of nongovernmental academic institutions, such as the Jameel Poverty Action Lab at MIT, which offered services of RCTs in international development. In 2006, the Center for Global Development published the Evaluation Gap Working Group's report on rigorous impact evaluation, which triggered a large-scale debate over the privileged use

of RCTs. The World Bank and other nongovernmental institutions increasingly valued RCTs as the strongest evaluation designs for assessing program impact. Finally, the U.S. Agency for International Development established their evaluation policy in 2011, promoting RCTs as the preferred approach for impact evaluations. One problem was that RCTs are not feasible for all policy topics, such as macroeconomic or environmental policy interventions, which affect the total population—as opposed to a discrete medical intervention.

Institutionalization of the RCT approach

As discussed in chapter two, the Federal Drug Administration had regulated that each new drug needed to be based on RCT evidence. Education and international development are far from this stage. However, these policy fields have developed their own systems of making RCTs the preferred methodological choice, and they have established evidence hierarchies and policy priorities.

Since 1970, the FDA requires that each new drug be supported by at least two positive RCTs in phase-three trials before it can be distributed in the market. Some exceptions apply to rare diseases and long-known drugs such as acetylsalicylic acid. Note that RCTs are only required in one stage of drug evaluation, and thus they are not the only recognized method in the larger evaluation process.

In the field of education, the U.S. Department of Education's (USDOE) priority for scientifically based evaluation in education (2005) is not an absolute priority. Although USDOE prioritizes RCTs over non-experimental methods when evaluating studies of similar quality, USDOE considers non-experimental research designs fundable if they demonstrate high scientific quality. The RCT movement in U.S. education has influenced the discourse and culture of educational evaluation through review institutions. One of them is DOE's What Works Clearinghouse, which functions as a review filter for RCT evaluations. One core problem was that WWC only found a few programs to be effective when employing the RCT evidence hierarchy. Therefore, school districts and education practitioners did not find the reviews useful for their curriculum planning (U.S. Government Accountability Office, July 2010).

In the arena of international development, institutions do not have the power of legislating methodological choices in evaluation approaches. Funding of development aid is diverse. However, private foundations such as the Gates Foundation or the Hewlett Foundation favor experimental evaluations; both foundations fund domestic and international projects. Furthermore, a few institutions have committed themselves to experimental and quasi-experimental evaluation designs. The World Bank established the Spanish Impact Evaluation Fund (SIEF), which has only been funding experimental and quasi-experimental evaluations. Originally, they restricted their funding to RCTs, although they opened the pool in the following year. It is unclear what triggered the change, but most likely the RCT criticism of certain constituents supported the change.

Influencing policy via funding choices rather than regulations on methodology is less obvious, but it can be equally powerful. For example, RCT funding priorities could increase funding for conditional cash transfer programs, but not for macroeconomic changes.

Countermovements against RCT primacy

In all three policy fields criticism and debates have arisen overt the value of RCTs to demonstrate effectiveness. In medicine, the physician Archibald Cochrane, known as the father of Evidence-Based Medicine (EBM), had already pointed out snags with the RCT methodology. He suggested being cautious when large sample sizes generated statistically significant results (Cochrane, 1971). Since the 1990s, the model of personalized medicine questioned the idea of the average patient, a notion upon which the RCT model was based. Instead, medical clinicians demanded personalized care tailored to the individual patient rather than a patient group. Sir Michael Rawlins, chair of the U.K. National Institute for Health and Clinical Excellence (NICE), remarked that RCTs had been put on an "undeserved pedestal" (Cuthbertson, 2008). Rawlins argued that RCTs have limited external validity because they are typically used for "specific types of patients for a relatively short period of time." Instead, Rawlins recommended a diversity of approaches that involve analyzing the totality of the evidence base, including observational approaches.

In U.S. education, the opponents of the RCT movement were mostly school practitioners and administrators, education evaluators, and theorists. They vehemently opposed the federal administration's demand for RCT primacy. Instead they argued for methodological pluralism. They were initially unsuccessful in getting their cause heard, as the AEA counterstatement to the proposed Federal Priority of experimental evaluations showed (cf., chapter three).

In international development, the Center for Global Development's 2006 report on the evaluation gap stirred disagreements from development evaluators and practitioners over the value of RCTs for evaluations of development interventions. The Network of Networks on Impact Evaluation (NONIE) directly responded to the CGD report and created a community of development practitioners and theorists who were interested in high-quality impact evaluations. NONIE members agreed that evaluation questions and contextual factors drive methodological choices, and not the reverse. There was disagreement about the value of RCTs. As shown in chapter four, some NONIE members were heavily opposed to the RCT primacy and instead promoted methodological pluralism or egalitarianism. Others insisted on the superior quality of the RCT approach, provided its implementation was feasible.

RCT critics across all three policy fields saw a threat of the RCT primacy in crowding out other valuable, and sometimes more appropriate evaluation methods. They were concerned that privileging RCTs would create narrow policy solutions for complex problems.

Adjustments within the RCT movement

In all three policy fields, the "RCT pendulum" is swinging away from RCT primacy back to the center. The FDA commissioner Margaret Hamburg called for more flexible standards of drug evaluations and for a reevaluation of methodologies beyond the RCT. She suggested revisiting unsuccessful RCTs in the past through the lens of biomarkers and turning them into positive findings for subpopulations of respondents with certain genetic markers.

The funding of comparative effectiveness research by U.S. legislators was an important step toward embracing a more inclusive evaluation approach in medicine. Up until now, the FDA has not required drug testers to use equivalency control groups, i.e., control groups receiving the best available drug on the market (cf., chapter two). However, findings from no-treatment controls are not as relevant for clinical practice, and thus they do not greatly inform decision making (i.e., doctors typically prescribe an FDA-approved drug for an ailment, but the question is whether there is a more effective drug than the one originally prescribed). However, in the area of education and international development, comparative effectiveness research has not yet been widely used. For example, in the STAR trial, the control groups could not utilize the same amount of resources for another intervention like extended school day or school year. Similarly, the Mexican Conditional Cash Transfer experiment used no-treatment conditions as controls. Small-scale experiments in Malawi and Morocco have so far incorporated comparative treatment arms with conditional and unconditional cash transfers (Banerjee & Duflo, 2011, 80). This would make comparative effectiveness evaluation possible and more relevant.

In the field of education, the U.S. Department of Education (DOE) did not respond initially to the criticism of the narrow definition of evidence, but later USDOE expanded the evidence hierarchy to include single case studies reviews (2010), as shown in chapter three. John Easton, the new director of the Institute of Education Sciences, also expressed a more inclusive view of methodological choices. He stated that "we must expand our repertoire of rigorous methodologies," which included non-experimental evidence (Easton, speech, March 28, 2011). Since the Obama Administration, a desire for more relevant, multi-methods evaluations of educational programs arose, swinging the methods "pendulum" to a more central position. At this point in history, it is unclear whether this desire has yet redirected education funding towards more inclusive evaluation approaches.

In international development, USAID's new evaluation policy (2011) underscored the value of both RCT designs and qualitative approaches. At first blush, they appeared to

fully integrate criticism against privileging just one method, and they argued for evaluations based on best methods à la NONIE's Alternative Group (U.S. Agency for International Development, January 2011). They further argued that evaluations should use methods that generate the highest quality and most credible evidence that corresponds to the questions being asked. Their statement promotes a well-balanced mix of qualitative and quantitative, experimental and non-experimental methods. Yet, their stance on the topic of methods for impact evaluation belied this openness: "Experimental methods generate the strongest evidence," and "alternative methods should be utilized only when random assignment strategies are infeasible" (U.S. Agency for International Development, January 2011). These statements do not align with their original quest for methodological balance, but treated the RCT as preferred method. But in general, USAID has been open to refining the question of methodological choice, and they are also considering alternative approaches.

In sum, the RCT movements in education and international development were inspired by the RCT use in medicine. They arose from a general search for a more scientific evaluation approach and from dissatisfaction with evaluation quality. While FDA directly institutionalized the RCT via requiring 2 RCTs per new drug approved, USDOE and USAID established preference priorities for RCTs by prioritizing funding of programs with an RCT attached. All three policy fields managed to federally regulate the preference for experimental designs in their way, but they also responded to criticisms and ultimately adjusted their evaluation policies accordingly. In the following section, I show that a closer look at RCT criticism in medicine can be helpful for soundly evaluating the RCT model as an evaluation tool for policy makers in education and international development.

2. Comparing arguments of RCT supporters and critics

The pendulum is in motion from RCT primacy toward greater methodological pluralism. The catalyst for this shift arose from the different views of RCT supporters and RCT critics. A closer look at these differences generates a more nuanced understanding of the RCT as evaluation tool. In the following section, I identify terminological and epistemological differences between the views, but I also show that core beliefs are

shared between the two sides. This shared base could become the basis for constructive discussion.

Terminological differences

Chapter four revealed terminological differences in the language use of RCT evaluators and their critics in international development. Howard White rightly argues that as long as these terminological differences prevail, a constructive debate about methodological choices can hardly take place (White, 2009). Evaluators have varied understandings of the notions of, for example, impact, counterfactual, attribution, and randomized controlled trial. These different understandings are typically due to the varied professional background of evaluators. Table 11 schematically compares a few terminological differences between RCT supporters and RCT critics.

TABLE 11: Terminological differences between RCT supporters and RCT critics

Sides / ***Term***	***as defined by RCT supporters***	***as defined by RCT critics***
(a) Impact	Net effect	Long-term effects
(b) Counterfactual	Theoretical concept	Real control group
(c)RCT	Statistical control for differences	Physical control

(a) Definition of impact

One crucial dividing line is how RCT supporters and RCT critics understand the term "impact." The divergence in definitions convolutes the RCT debate. When using the term "impact" in development evaluation, both RCT supporters and RCT critics typically refer to the OECD DAC's definition, but emphasize different parts. RCT critics underscore multidimensional aspects, such as positive and negative, primary and secondary, direct and indirect, intended and unintended effects; as well as economic, socio-cultural, institutional, environmental, and technological changes (cf., OECD Development Assistance Committee, 2010), none of which an RCT would be able to easily measure. One of many challenges in measuring such multidimensional aspects would be the need for the unit of assignment and analysis to coincide. In contrast, RCT supporters take a more narrow focus on impact as "net effect," i.e., comparing the outcomes of an intervention with what would have happened without the intervention (cf., Gertler,

Martinez, Premand, et al., 2011). They typically rely on a limited number of measures and then compare the baseline with the end measurements. Unanticipated effects, by definition, would not be captured by baseline measurements and would thus not be part of an RCT evaluation. RCT critics rightly point out the discrepancy between what RCTs typically measure and how multidimensional impact is defined.

I have already shown how the supporters and critics of RCTs emphasize different aspects of the definition of "impact." Whereas RCT supporters tend to emphasize the causal aspect, RCT critics typically emphasize the multi-dimensional aspect of the definition. Both aspects are part of the definition, yet shifting the focus slightly has major consequences for methodological ideologies. This dissonance highlights how difficult it is to standardize a term's use in the sub-fields of evaluation, where users' background and training casts different meanings on methodological concepts. The misunderstandings that have resulted underscore the need to agree on terminological meaning.

(b) Definition of counterfactual

RCT supporters and RCT critics use the term "counterfactual" differently. RCT supporters do not necessarily think that a randomized control group produces a real counterfactual. They think of the counterfactual as a theoretical idea, which a well-controlled comparison group would merely approximate. In fact, the randomized control group could never simulate a true counterfactual. Counterfactuals by definition can not exist, because they refer to what would have happened if a person or group had not received an intervention. Counterfactuals might be approximated by randomly assigning different subjects to different interventions. RCT supporters typically believe that determining the impact—i.e., net effect—requires a well-controlled comparison group, which would imperfectly represent the counterfactual via randomization.

RCT critics, however, tend to equate the counterfactual with a randomized control group. They do not believe in the necessity of a so-called "counterfactual analysis" to establish causal impact. So when both sides seem to disagree on whether a counterfactual is

necessary in determining program impact, they may not be talking about the same concept.

The different understandings of counterfactual lead to the question of to what degree an impact evaluation needs to approximate a counterfactual situation. RCT supporters and RCT critics seem to disagree in this regard. Obviously, a true counterfactual is not necessary for knowledge generation because it does not exist by definition. In general, a counterfactual situation, if simulated either by a control group or statistically, would help in determining impact. However, a factual analysis of the program (without a comparison group) might also be sufficient in determining impact, similar to knowledge generation in the natural sciences. This would be especially true when there are few external factors that could influence impact. Howard White, for instance, referred to an intervention for making water transport more efficient. An impact evaluation of such an intervention did not require a counterfactual because no additional changes were made to the water system (White, Barbu, 2006).

(c) Definition of randomized controlled trial

In general, experimental and qualitative evaluators interpret the concept of the randomized controlled trial itself differently. The common rationale for performing an RCT is the idea of establishing a causal connection through the use of a simulated counterfactual, represented by the randomly assigned control group. For experimentalists, the RCT stands out from other approaches through the concept of random assignment and the existence of a (randomly assigned) control group. In the tradition of Ronald Fisher, RCT supporters emphasize that these two characteristics make experiments suitable for the field beyond an artificial laboratory.

Evaluations can take place in a laboratory setting or in a field setting. Wilhelm Wundt founded experimental psychology in the 19th century and created the experimental laboratory wherein researchers tested humans in a controlled environment. Its purpose was to reduce the variance of possible influential variables. Research subjects were isolated from the routine of ordinary living. Researchers then manipulated the variable under study under rigorously specified and controlled conditions (Kerlinger, 1966, 379).

RCT critics sometimes equate RCTs with these laboratory experiments. This is a false equation, based on a misunderstanding of the nature of RCTs. This misunderstanding might arise from an evaluator's limited knowledge of quantitative methods, and of RCTs in particular. RCT critics sometimes regard the terms "RCT" and "quantitative" interchangeably, which is also incorrect. In fact, RCTs have many qualitative elements, such as choosing sample and interpreting findings (cf., policy recommendations).

RCT critics sometimes believe that the term "controlled" in randomized controlled trials refers to the externally controlled context of the treatment and control group, as would be the case in a laboratory experiment. They sometimes think that RCTs exercise control over the study subjects by controlling the subjects' environment. However, the nature of control lies primarily in the random assignment process, not in the control of exogenous and endogenous variables. Therefore, the RCT would actually allow for minimal environmental control because it *statistically* controls exogenous factors via randomization. Fisher was right in his observation that RCTs equalize the innumerable factors and causes of disturbances that would change soil fertility (cf., chapter one, Fisher, 1935, 21). The theoretical equalization of exogenous factors allowed researchers to conduct fertilizer studies outside the laboratory in the first place. RCT supporters emphasize this advantage of requiring less control. Still, RCT critics are right in that RCTs are controlled in the sense that RCTs typically follow a strict protocol with obtaining baseline and end measurements, in monitoring and managing attrition in both study groups, and in making sure the intervention is implemented correctly.

RCT supporters and RCT critics would benefit from making their terminology even more transparent, especially given that the evaluation field encompasses so many professional backgrounds. The OECD DAC glossary on key terms in evaluation and results-based management is a successful example of how to achieve definitional transparency (OECD Development Assistance Committee, 2002), but even a glossary leaves room for interpretation as the term "impact" revealed. I suggest that a similar glossary of impact evaluation would be a worthwhile starting point for bringing clarity into the controversy over RCT's role in evaluation. The NONIE guidelines, as discussed in chapter four, were

intended as a start, but they did not result in the anticipated common ground for defining and guiding impact evaluations. A renewed effort is needed to reach common ground.

Epistemological differences between RCT supporters and RCT critics

Apart from terminological differences, RCT supporters and critics exhibit epistemological differences, i.e., differences stemming from how humans think about the world. Applied to the field of policy evaluation, epistemological differences arise from how RCT supporters and RCT critics think differently about how evaluation generates knowledge. Thinking about these differences allows a more nuanced understanding of the RCT as a policy tool. RCT supporters *versus* RCT critics would disagree in the following topics, organized in Table 12 :

TABLE 12: Epistemological differences RCT supporters and RCT critics

Sides / ***Term***	***RCT supporters***	***RCT critics***
(a) Hierarchy	RCT as apex in methods hierarchy	No evidence hierarchy, but methodological equality
(b) Sources of evidence	Evidence from (quasi-) experiments only	Evidence from experimental and non-experimental approaches
(c) Complexity	RCTs are preferred for complex interventions in complex situations	RCTs are not suitable for complex interventions in complex situations
(d) Concept of science	Universal rules	Changing rules
(e) Participatory evaluation	Evaluations need to be scientific in the first place	Evaluations need to be participatory
(f) Best method	Yes, should be used	Yes, should be used

(a) Hierarchy: RCT as apex in the methods hierarchy?

One fundamental question is whether the RCT should be the apex in the methods hierarchy to establish causal impact. For the RCT supporters, the answer is yes. The first question an evaluator should ask is whether an RCT is feasible (cf., Subgroup 1's NONIE document). Whether an RCT is feasible depends, among other things, on the number of units of assignment (White, 2010, 155). For small numbers of assignable units, some RCT purists would argue that no causal statements about program impact can be made because an RCT is not feasible. For RCT pragmatists, quasi-experimental and other

quantitative approaches could be used to attribute causal effects to an intervention with few or just one assignment unit. For example, a macroeconomic intervention to adjust inflation or interest rates could rely on historical data to demonstrate impact. However, such evidence would not be as strong as the evidence produced by an RCT, because there would not be a strong counterfactual like the randomized control group.

For qualitative evaluators, there is no hierarchy of methods. The EES statement, for example, emphasized the equality of methodological approaches (cf., chapter 4). Therefore, the RCT should not be the default approach for determining causal effects. They would further argue that non-experimental, observational evidence is equally valid for determining causal effect. They refer to those areas of the natural sciences and humanities that rarely use RCTs to arrive at their answers. These areas do not treat non-experimental approaches as second class.

Even Sir Michael Rawlins, chair of NICE, suggested moving away from the vertical concept of evidence assessment (cf., chapter two), where there would be a clear winner at the apex (i.e., the RCT). The question is what should be put in place of a hierarchy. Referring to the equality of approaches does not give the evaluator and policy maker any guidance for when to use what method.

(b) Sources of evidence: From (quasi-)experiments only?

RCT supporters and RCT critics typically disagree as to whether non-experimental methods can generate evidence on causal effect. As previously indicated, non-experimental approaches such as interpretive data analysis do not necessarily use an explicit counterfactual, but are factual in nature instead. Qualitative analysis is focused on what has actually happened instead of what would have happened without the treatment.

RCT evaluators argue that an impact evaluation requires at least an implicit counterfactual, and that the RCT produces the strongest version of such a counterfactual. Qualitative evaluators, however, may not think that a counterfactual is necessary to determine impact. They tend to regard factual analysis as sufficient for determining causality citing natural science and historical research (Scriven, 2008).

The "smoking and lung cancer" correlation is a commonly cited example used by experimentalists and qualitative evaluators to make a case for experimental or non-experimental analysis (cf., chapter three). Both sides agree that RCTs cannot be ethically implemented in the smoking case. The fundamental causal question is whether non-experimental evidence suffices for establishing that smoking causes lung cancer. Non-experimental evidence could be quasi-experimental, quantitative and non-experimental, or qualitative. In the lung cancer example, experimentalists might emphasize that randomized smoking experiments were done with animals. With animal experiments, however, the question of external validity has not been resolved. Because RCTs were only performed on exemplars of animal species, the applicability of their findings to the human species is unclear. They might argue that quasi-experiments in the human species have shed further light on the connection, where treatment and control persons were matched by individual characteristics. However, as with all quasi-experimental designs, the question of "unobservables" is unresolved. Therefore, experimentalists would argue that the evidence about smoking and lung cancer is not as strong as it would be with an RCT. Conversely, qualitative evaluators would argue that non-experimental analysis of smoking and lung cancer is sufficient to determine causal effect. In particular, systematic observation and measurement of individual cases could transfer the correlation into a causal connection by excluding alternative explanations (Scriven, 2008).

The "smoking and lung cancer" case ties into the previous issue of methods hierarchy and illustrates the need for a more inclusive understanding of evidence, where non-experimental sources are also used.

(c) Complexity: RCTs suitable for complex interventions in complex situations?

RCT supporters and RCT critics disagree as to whether RCTs can adequately evaluate the effects of complex interventions in complex situations. Patricia Rogers and others have argued that RCTs are appropriate for simple linear interventions, such as drug treatments or the provision of educational materials, while pointing out that RCTs are inappropriate for complex interventions with multiple determinants (cf., chapter four). Such complex interventions in complex environments include sector-wide approaches at a country or

regional level—such as macroeconomic interventions that change interest rates, espouse human rights intervention, or lead to institutional changes—and interventions continuously adapted to a changing environment.

However, in his article on the summative evaluation of the RCT approach, Scriven recognized the usefulness of RCTs, especially in complex situations where evaluators would have a difficult time accounting for all influencing factors (Scriven, 2008). In the 1920s, Ronald Fisher had already stated that one major reason to conduct RCT studies was to randomly equalize the heterogeneity of the environment (Fisher, 1921). Under this lens, RCTs seem very useful for determining effects in complex situations. The main problem is that RCTs per se are not able to answer the question of why the interventions were effective in such complex situations. Therefore, RCT findings cannot be readily applied to other settings.

(d) Concept of science: Universal rules or changing rules?

The fundamental difference between RCT supporters and RCT critics is their view of what science is and what it can do. NONIE members consulted Patton when they put together their draft on alternative methods for causal impact evaluation (cf., chapter five). In the newest edition of his *Utilization-Focused Evaluation* (2008), Patton added a section on why RCTs should not be the gold standard in evaluation. Patton criticized experimental evaluation as being a positivistic enterprise that assumed an objective reality based on immutable laws and universal rules. Evaluation would therefore become a science of "just getting the facts right," and would profess to describe things as they really were and how they really worked—which Patton criticized as an illusion (Patton, 2008, 8). Patton countered that in science bare facts did not exist. Whether things would work depended on the sociopolitical and cultural context.

Anthony Bryk rightfully pointed out that the question was not whether "something works" in general and whether there was some "fixed, true, general effect" to be estimated. On the contrary, the question was what effects accrue when enacted by different individuals working under varied contexts (Bryk, email communication, January 2, 2011). He continued: "Rather than a fixed treatment effect, intervention effects exist in

a multi-facet[ed], multivariate distribution. Consequently, depending on where you inquire, different results will occur" (ibid.). These variable intervention effects in various contexts make it difficult to insist on such general rules. This does not mean that the RCT is not useful in knowledge generation. Patton agreed that the ideal way to control extraneous influences was to randomly assign units to treatment and control groups (Patton, 2008, 447). The RCT would create a "hypothetical counterfactual" of what would have happened to the treatment group if they had not received the treatment. Random assignment would distribute known and unknown extraneous causes evenly across treatment and comparison groups, and therefore average the effect of extraneous factors. The RCT, however, does not provide a window to universal rules, but rather to changing rules, depending on the particular context.

(e) Participatory evaluation?

RCT supporters and RCT critics differently value the importance of political inclusion and participation in an evaluation. RCT supporters tend to focus on issues of selection bias and internal validity—that is, whether units were selected by chance and whether the program intervention had produced the desired net impact. Usually outside entities implement RCTs in program beneficiary groups. Because RCTs require random assignment by an unbiased entity, they have the reputation of being "guinea pig" experiments, in which people are treated as study objects rather than human subjects. Humans or households cannot decide whether they want to be in the treatment or control group.

RCT critics, such as members of the NONIE Alternative Group, are interested in the political process of evaluation and are concerned about the appropriate inclusion of program stakeholders in the evaluation process (cf., chapter four). They fear that RCT implementation disempowers intended program beneficiaries by having external evaluators assess the program. They prefer to employ local evaluators with local knowledge, who would be less likely to suggest randomized allocation. RCT critics see RCTs as a Western invention that is imposed on the developing world (Duncan, 2008). The evaluation theorist Jennifer Greene even accused RCTs of being "inherently undemocratic" because data are extracted from the local population without engaging

them (Greene, presentation, April 2, 2009). The evaluation practitioner Zenda Ofir, former president of the African Evaluation Association, found that the RCT movement itself was a way to control and exert influence over countries, along with war, trade policies, investments, and other strategies (Ofir, email communication, August 6, 2007). Representatives from developing countries felt that the RCT approach was foreign to their way of thinking, and they thus perceived it as a new type of colonialism.

However, RCTs do not necessarily need to be undemocratic or anti-participatory, and RCT supporters would deny criticism implying that they are either. The economist Esther Duflo prided herself in spending a great deal of time "speaking with people in the villages, not sitting in the capital city talking to donors" (MIT Technology Review, January 2010). Speaking with people is a start, but does not yet make an evaluation participatory. It is still the case that representatives of Western countries initiate and conduct RCTs in developing countries. Even in the Progresa case, the Mexican decision makers had received their training in the United States, so they were indirectly influenced by Western thought. The World Bank and other multilateral agencies provide capacity building to evaluators in developing countries (e.g., the Africa Impact Evaluation Initiative), but even then knowledge transfer emphasizes the original knowledge gap.

(f) Best method: Should the best method be used?

RCT supporters and RCT critics might agree that an evaluator should use the best method available (White, 2010, 162). However, they are polarized about what the best available method would be in certain instances. For RCT supporters, the decision tree for methods choices always starts with the question: Is an RCT feasible (cf., Figure 4)? Experimentalists do prefer RCTs for causal questions with large sample size; if randomization is not possible due to ethical or political constraints, such as in the smoking-cancer example, they suggest using other quantitative quasi-experimental designs. For qualitative evaluators, no such decision tree exists. They do not give RCTs this preferential treatment. Instead, they believe in methodological pluralism and methods equality without a rigid methods hierarchy. In his book, *Developmental Evaluation* (2010), Patton found the real gold standard for impact evaluation methods to be "methodological appropriateness"—that is, matching methods to the nature of the

question and the purpose of the evaluation, rather than "blind adherence to one particular design" (Patton, 2010, 39). What methodological appropriateness means is a question still up for discussion. Qualitative evaluators have not explicitly demonstrated how they go about deciding which method to use.

In the field of evaluation, experts often describe themselves as either qualitative or quantitative evaluators. This distinction signals their interests and expertise. A qualitative evaluator is more likely to use participatory and utilization-focused approaches, while a quantitative evaluator might have a strong background in experimental and quasi-experimental designs. Their segregation into qualitative, quantitative, or experimental evaluators does not mean that there are insurmountable differences in their views, but degrees of disagreements.

RCT supporters value the strengths of experimental designs more than qualitative evaluators do. They believe that an impact evaluation needs to establish internal validity before questions of external validity can be raised (Glennerster, presentation, November 10, 2008). They maintain that qualitative methods themselves are not able to establish internal validity, although these might assist in strengthening external validity. In contrast, qualitative evaluators may emphasize the strengths of qualitative tools and when and where they would be superior to quantitative approaches. They also like to point out challenges of RCTs, which RCT supporters often know equally well, but may find less problematic because they focus on internal validity as an RCT's strength.

There is a range of viewpoints between the extremes of "RCT primacy" and "methodological pluralism." There are many RCT supporters and RCT critics who do not adhere to either extreme of the continuum, but who stand more in the middle. They might be able to have a constructive discourse about their perspectives on methodological choice.

Although this section has so far emphasized the divergent viewpoints, RCT supporters and RCT critics have a common ground where a constructive dialogue could and should start. First, both sides agree that an impact evaluation should be driven by the particular

evaluation question and not by any preferred method. Second, they both agree that a triangulation of different methodological approaches and tools would increase the quality and relevance of an impact evaluation. Third, they acknowledge challenges of RCTs across policy fields. These three points of agreement could stimulate a constructive conversation between RCT supporters and RCT critics toward broadening the methodological toolbox.

3. Challenges of RCTs

The RCT theorist Fisher argued that a statistician's task was to determine how to evaluate the limitations of the data in hand and to recognize the limitations of the experimental approach (Fisher, 1933, 46). This insight also holds true for decision makers and policy makers who are the audience for RCT results. Often times, however, they focus on the strengths of RCTs (i.e., internal validity and scientific authority), while forgetting to also point to the limitations and challenges of RCTs.

TABLE 13: Challenges of RCTs by topics in medicine, education, and international development

Field / ***Topic***	*Medicine*	*Education*	*International Development*
(a) Heterogeneity of target variables	Heterogeneity in target patients	Heterogeneity in target students, teachers, education systems	Heterogeneity in target cultures, regions
(b) Heterogeneity of study population	Heterogeneous study patients	Heterogeneous study participants	Heterogeneous study participants
(c) Black box of the intervention process	Pharmacodynamics and pharmaco-kinetics	Learning and teaching process	Socio-cultural process
(d) Insignificant findings	Due to wrong dosage or administration	Due to missing catalysts such as experienced teachers	Due to missing catalysts such as proximate health clinics
(e) Capture long-term impact	Mortality and quality of life	Increased income	Intergenerational poverty reduction
(f) Political process	Drug manufacturers	Textbook companies	Development entrepreneurs
(g) Change of policy priorities	Frequent diseases	Phonics instruction	Health interventions

The policy fields of medicine, education, and international development are each faced with their own nuanced group of limitations and challenges. Understanding these challenges will help evaluators and policy makers value the RCT in a realistic way. TABLE 13 summarizes challenges of using RCTs in policymaking by topic. I explicate these based on materials from previous chapters.

(a) Heterogeneity of target variables

One shortcoming of RCTs is the limited applicability of findings to other populations and contexts, especially when the RCT is locally restricted and confined to a select pilot sample. First and foremost, RCTs target the question of causal attribution whether a specific intervention in a particular context brought about certain outcomes (Pawson & Tilley, 1997). RCTs do not automatically support the transfer of particular findings to other populations and contexts.

In the field of medicine, the heterogeneity of patients outside a trial has been a major concern. Often times, the patient sample does not represent the general run of patients, as Austin Bradford Hill pointed out (Hill, 1937, 9). The exclusion and inclusion criteria, as outlined in the clinical trial registers, prevent many patients with the medical condition under study from participating in a trial. For example, RCTs often exclude comorbid patients who suffer from more than one disease. Therefore, trial findings do not automatically apply to these patients.

In education, the heterogeneity of students, teachers, and the education system in general poses a challenge. The Tennessee Student-Teacher Achievement Ratio experiment, for example, found that smaller class sizes increased academic student performance in the pilot (cf., chapter three). However, the Tennessee evidence did not map directly onto California, a state that serves a much larger and more diverse student population than Tennessee does. Furthermore, class sizes in the Californian school system were originally much larger than in Tennessee, and thus, while the reduction was of similar size, the reduced class size was still larger than the class size in Tennessee. The Tennessee RCT only answered the question of the reduction of class size from 24 to 16 students, but it did not address what would happen if class sizes were reduced from 30 to 20. Finally, the

Tennessee trial did not address the larger-scale issue of teacher supply. The state of California, for instance, did not have the required supply of teachers to appropriately implement class-size reduction. This example illustrates the difficulty of transferring evaluation findings to other contexts.

In international development, applying RCT findings across different national and cultural contexts is even more problematic due to their heterogeneity. The application of RCT findings may turn out differently in culturally and economically distinct societies. The problem of heterogeneous target populations in international development can be illustrated by the RCT evaluation of Progresa, a conditional Cash Transfer Program in Mexico. Many contextual variables existed that influenced the success of the intervention. Were the program applied to fathers or in an Arabic country, the results might have been different.

(b) Heterogeneity of study population

A related problem is representativeness of findings for heterogeneous study groups. RCT findings foremost yield an average impact estimate on the particular study population (e.g., an average human or an average household). RCTs, however, do not provide individual-level impact estimates. Such average humans and households are artificial constructs and thus are not meaningful for policy makers who are concerned about individual-level policy solutions.

In medicine, average findings of drug effectiveness may not be relevant for patients with certain heterogeneous characteristics. In education, average findings of a reading intervention may not be relevant for a bilingual student. In international development, conditional cash transfers may be less effective for households living far from a school building. Estimating impacts for subgroups, such as elderly patients, would be more informative for decision makers. However, RCT experts rightly believe that subgroup analysis is not permitted in an experimental design when random assignment has not taken place within those subgroups in the first place. It is often hard to predict which subgroups will exhibit unexpected results. This shortcoming does not allow evaluators to determine how the program affected individuals or subgroups. Policy makers, however,

would like to know more than average and size of the measured effect because they must address needs of various constituents. For example, a legislator might not want to promote an education initiative that would increase the achievement gap for minority students, or he might not feel comfortable suggesting a certain job placement program that would put African-American women at a disadvantage. An RCT is typically not designed to answer questions specific to subgroups. If it were, sample size would need to increase significantly for those subgroups in order to achieve statistically meaningful results.

(c) Black box of the underlying intervention process

RCTs are black box evaluations (Bourguignon & White, 2007). The term "black box" refers to the idea of a machine that receives inputs and generates outputs. The process in the box is not visible to outside observers. For example, psychologists characterize human consciousness as a black box, whose underlying structure, dynamics, and mechanisms are not fully understood (Silverman, 2006).

In drug trials, black box evaluations are common. The mechanisms of many psychotropic drugs are not known. For example, while Bupropion was tested as a smoking cessation drug, researchers observed its mood-lifting capability. They tested the drug for its antidepressant effects. Although researchers confirmed the effects, they were not able to determine the exact mechanism for how this drug worked in the human body, also called pharmacokinetics. For medical purposes, opening the drug's black box is not necessary because its active ingredient, i.e., Bupropion, is known. So the drug can be repeatedly manufactured using this ingredient. This is typically not the case in non-medical intervention. A program often consists of several components and human mediators. For example, a new education program does not just consist of a textbook and manipulatives, but also requires teachers to implement the program and students to collaborate. That is to say, an education program does not simply work like a medication that a person swallows; an education program operates in a complex social and cultural environment. Knowing the pathways of the impact would be helpful in applying RCT findings to other policy contexts. The black box phenomenon is even larger with complex and multidimensional programs, because more factors interact. Patricia Rogers pointed out

that evaluators would need to carefully develop a program theory to start opening this black box for complex programs (Rogers, 2008).

In medicine, preclinical testing in laboratories also serves the purpose of understanding the biological pathways through which a drug works. In the case of education and international development, such laboratory testing would be difficult to perform.

To a certain degree, RCTs could assist in opening the black box, according to J-PAL associate Bruno Crepon (Crepon, personal communication, June 18, 2008). Crepon argued that one could construct treatment and control groups that test the assumptions of those underlying processes. For example, a job creation program for hard-to-place individuals could provide these individuals with new skills (i.e., the treatment group) or could give employers incentives to hire these individuals (i.e., the control group). The comparison between those groups could test whether job creation programs work through the mechanism of skills or employer incentives. A major drawback of this experimental process is that many RCTs would need to be performed to really determine the precise mechanism of an intervention. Evaluators acknowledge, therefore, that RCTs alone do not open the black box of how a program works, and they understand the importance of qualitative-interpretive approaches for opening this black box. For example, qualitative-interpretative observations, similar to the role of preclinical testing in drug development, could help in investigating the underlying processes and channels through which an intervention is effective.

Qualitative-interpretive studies generate a different kind of evidence via observations and interviewing, which allow for establishing a hypothesis and then following iterative data interpretation. Although experimentalists are skeptical about using interpretive data analysis for determining causal effect, they would agree that it could explain causal processes.

(d) RCT evaluations often produce insignificant findings

Still, a major challenge of RCTs has been that the majority of findings are statistically insignificant. Often, the large-scale federally funded RCTs in the 1980s arrived at

statistically insignificant results. Thomas Cook found that most educational RCTs yielded statistically insignificant findings (Cook, 2006). As a result, policy makers became frustrated over the fact that expensive federal experiments did not yield the desired results and thus eventually reduced funding. From a policy perspective, there is a danger that policy makers dismiss interventions based on insignificant findings from RCTs, given that statistical insignificance does not necessarily imply overall program ineffectiveness. Regarding statistical significance testing, recall that Ronald A. Fisher regarded the level of probability of .05 as merely a convenient convention (cf., chapter 1). Today it is often applied as a rigid standard for deciding whether a program works. RCT results with a *p* value of less than .05 are then automatically called insignificant findings, but maybe not rightfully so.

Findings may appear to be insignificant for several reasons. One is that, indeed, the intervention is ineffective. However, another explanation is that the intervention in its current form or in its current context is on average ineffective; but small changes in the study population, the study context, or the intervention itself might yield significant findings. This potential misrepresentation of an intervention's effectiveness leads to a major problem of RCTs: they only test a small portion of possible policy interventions in possible contexts.

Statistical insignificance can also result when the tested intervention is a necessary but insufficient component to produce statistically significant results. Lack of treatment integrity and failures in implementing the program may exist. The odds of underestimating effects in RCTs are larger than the odds of overestimation (Wittmann, 2011). Medical clinicians are often disappointed that the interventions they use in daily practice do not prove to be effective in RCT evaluations (Wittmann, Klumb, 2006). As noted above, statistical insignificance does not automatically imply program ineffectiveness. Other catalysts may be needed to make an intervention effective. One example comes from an RCT suggesting that textbooks proved ineffective in increasing student performance (Kremer 2008). What this RCT did not find out was whether teachers had the necessary training to use the textbooks effectively, or whether the textbooks were suitable for the particular students. Teachers might have needed

additional training in using these textbooks effectively, or the books' content or language might not have been age appropriate. Many reasons may exist for why the particular textbooks were not an effective way of instruction in this particular context. Only qualitative inquiry would help answer the question of *why* textbooks were ineffective and *how* to improve their effectiveness.

Statistically insignificant findings can be the starting point of a program's elimination. It is therefore necessary to ask whether the public program fails due to inadequacies of the program or due to deficits in the research approach used (Wittmann, 2011). Peter Rossi's Iron Law of social program evaluation pointed out that "nothing works" for many programs: "Most programs, when properly evaluated, turn out to be ineffective or at best marginally accomplishing their set aims" (Rossi, 1978, 574). Why would this happen? Over-reliance of significance testing may be one answer. The choice of inefficient evaluation approaches is another possibility. It is important to tailor evaluation methodology to practical demands.

(e) Capture long-term impact

RCTs are used to determine whether a program has had the desired impact. RCT critics refer to the OECD DAC definition of impact, which includes more aspects than just the desired impact (OECD Development Assistance Committee, 2002). These aspects include positive and negative, short-term and long-term, direct and indirect, as well as intended and unintended components of impact. The definition also encompasses economic, socio-cultural, institutional, environmental, and technological levels of effect.

Evaluations in general may suffer from a narrow time horizon with short-term impact measures, because funders may need a quick turnaround of one or two years to produce evaluation results (Bamberger, 2006). Frequently drug studies entail only a few weeks of administering a drug, even with chronical-use drugs such as cholesterol-reducing medication. Long-term effects, especially their side effects, cannot be studied in so short a time. In international development, Progresa is an example of a program that used a pipeline approach where the comparison group received the same treatment 20 months after the original treatment group did. The primary outcome measure was school

participation, which did not automatically predict whether students would complete their education and benefit from it in the labor market. Improvement of educational outcomes, increase of employment, and a general increase in income to break the generational cycle of poverty would be true impact measures, but these were difficult to measure. Ideally, Progresa would have excluded the control group from the intervention for several years, but that was not feasible, ethically or politically. Increasing the time horizon would have raised ethical issues of excluding the control group from conditional cash transfer once data on promising effects of Progresa became available.

When employing a comprehensive definition of impact, one single pre- and post-intervention measure of the outcome variable of interest is often insufficient. Wittmann and colleagues argued for longitudinal data analysis with several points in time that measure outcomes (Wittmann, Klumb, 2006). This allows evaluators to capture a "larger slice" of intervention effects over longer periods.

Furthermore, an intervention typically does not just affect one level, such as the household level. Policy interventions typically affect several levels in a society: individuals, households, communities, institutions, regions, and nations. RCTs cannot determine the causal effects at all of these levels. If an evaluation would like to comprehensively capture the intervention impact, it would need to not just measure the effects at the level where the original randomization took place, but also capture institutional, societal, and regional changes. Technically, the level of analysis must follow the level of random assignment. When communities are the unit of randomization, then evaluators should not analyze individual-level outcomes to determine program impact. Internal validity cannot be guaranteed, and the evidence is less strong. The same is the case for extrapolating findings to higher levels based on the original scheme of random assignment. The Guatemalan village study illustrated the temptation to perform individual-level analysis despite the fact that villages had been randomized (cf., chapter four).

(f) RCTs take part in the political process

The limitations and shortcomings that affect quantitative approaches in general naturally apply to RCTs. They may result from the political process to which evaluations belong. Time constraints and resource constraints often do not allow for high-quality experiments. As mentioned above, some quantitative measures used in RCTs might be narrow. The U.S. Head Start evaluation, for example, used test scores as outcome measures. The experiment did not use other less quantifiable outcomes, such as student motivation or self-confidence. Qualitative-observational components such as classroom observations might have captured important but less quantifiable results. At least in theory, some of these limitations could be addressed by changing funding cycles and donor requirements. These solutions apply for both quantitative and qualitative evaluations. To measure long-term effects and sustainability of programs, for example, longer time horizons are needed.

Donald Campbell argued that "the more any quantitative social indicator is used for social decision-making, the more subject it will be to distort and corrupt the social processes it is intended to monitor" (Campbell, 1976, 49). This statement is especially true for the use of RCTs in the political process, and it particularly applies when there are commercial interests involved. In the field of medicine, the Federal Drug Administration attempted to guarantee drug safety and ensure drug effectiveness by requiring that each new drug provide evidence from two RCTs. Researchers found, however, that if commercial drug manufacturers sponsor the RCT, the drugs are much more likely to show effectiveness than with non-commercial funders. Commercial bias is possible via opportunistic choices in the RCT process, for example, by using narrow inclusion criteria.

In education, there is some commercial bias as well. Textbook markets are worth multibillion U.S. dollars. Textbook companies jumped on the "bandwagon" of RCT evidence in the aftermath of Reading First and NCLB. Their interest was to show that their educational programs met the criterion of scientifically based evaluation. Rather than performing their own RCT evaluations, they initially aligned their products with

what was considered scientifically based. They ultimately crowded out products that did not fit the criteria.

In international development, certain "pet" programs have emerged. The CCT program is sold as hard evidence, although it is still unclear whether conditions are necessary. School-based deworming is another example where private foundations and governments were enthusiastic followers, but where RCT results did not add much knowledge.

(g) RCT findings change the policy focus

From a policy perspective, the new focus on RCTs might change how resources are allocated in the public sector. Programs in which RCTs can show positive effects might be funded in the future, whereas programs in which RCTs are not feasible or desirable might be discontinued. Under the federal Reading First initiative, reading programs that did not show an alignment with the evidence-based five reading pillars were excluded from funding. Phonemics programs for which more experimental evaluations existed were more frequently funded under the Reading First component of the 2001 NCLB, as illustrated in chapter three. Later, the What Works Clearinghouse (WWC) excluded observational non-randomized studies from their database. In international development, private foundations such as the Gates Foundation or the Hewlett Foundation have preferred to fund health and education programs with a strong RCT base.

There is a danger that policies may get evaluated and funded solely based on the fact that they are amenable to RCT evaluations. For example, RCTs are not possible for interventions that do not directly affect individuals or households, which means that their effects cannot be measured on an individual or household level. This is the case for macroeconomic, structural, and often environmental policies. The preference for RCT evaluations creates a project selection bias. Dina Pomeranz welcomed a shift in structuring aid delivery based on evidence from RCTs (Pomeranz, Neue Zürcher Zeitung, May 13, 2008). In her view, funding should be made conditional on evidence created by RCTs, i.e., only programs with an experimental basis would deserve and receive funding. But only funding programs with an RCT basis introduces bias for certain policy fields and certain policy responses within those fields.

In medicine, bias would be toward frequent diseases that have large income potential. RCTs cannot be performed for rare diseases, which also promise less income. In the area of literacy education, phonics instruction is more susceptible to RCTs than the whole of language arts education. In international development, health and education interventions are most easily measured by RCTs. Macroeconomic or environmental policies are less amenable to RCTs and they would thus potentially get less funding.

In this section, I have illustrated various challenges the RCT model faces. While all these challenges of using the RCT model are already present in the field of medicine, education and international development exacerbate these obstacles due their complex interventions and contexts. When RCTs are taken out of the context of treating an individual for a biological disease and into the context of treating student groups for educational performance or households for socio-economic problems, the challenges of applying the RCT model increase. Because the explanatory power of an RCT is always zero (thus "blackbox evaluations"), evaluators need to find ways of explaining the impact. This is where hermeneutic-interpretive observations—i.e., understanding the issue at stake—begin.

4. Policy recommendations for judging the RCT model

In order to use credible evidence of program impact, policy makers need to be alerted to several issues. At times, evaluators point to the strengths of certain methodologies, but they do not equally emphasize their limitations and where other methodologies should be preferred or at least added. This problem is especially true for the RCT model because it has attached an aura of "scientificness" and objectivity, no doubt partially due to its success in finding life-saving medicines (c.f., Tuberculosis and Poliomyelitis trials).

In the following, I provide five policy recommendations derived from the described RCT movements and debates. They are meant to inform how to evaluate impact in public policy (cf., TABLE 14). Some of these recommendations stem from the use of RCTs in the medical field, but can be also applied to the other policy fields of education and international development.

TABLE 14: Policy recommendations for evaluating impact

(a) Understand biases associated with RCTs
(b) Understand the qualitative-interpretive components of RCTs
(c) Clarify representativeness of RCT findings
(d) Expand evidence base beyond the RCTs
(e) Produce evaluations on comparative effectiveness
(f) Strengthen the evaluation function

(a) Understand biases associated with RCTs

Although in general RCTs minimize selection bias, they are not necessarily immune to other types of biases. Ernest House argued that, despite the obvious advantages of RCTs for establishing internal validity, they might be biased in other respects (cf., chapter 2; House, 2008, 416). House identified 14 sources of bias for RCTS in drug testing, resulting from opportunistic choices, such as using a homogeneous population (e.g., patients with only one disease). These biases occur to an equal extent in other policy areas, such as education and international development. The choice of a sample with homogeneous characteristics—e.g., students above the poverty level; households that have access to certain public facilities—can influence results. The choice of time scale can likewise influence results—e.g., when chronic-use drugs are tested for a short period or a reading intervention is tested only for a few weeks. These sources of bias do not pose threats to internal validity, but they negatively affect external validity, i.e., the representativeness of findings for other populations.

A conscious use of qualitative-interpretive approaches could help identify potential biases in RCT studies and determine to what degree RCT findings could be applied to other contexts. Apart from these opportunistic biases, one needs to keep in mind that no research approach, including the RCT with its hallowed association with the sciences, is able to produce objective information. An RCT result is not a bare fact. RCTs generate a number that quantifies the difference between treatment and control group measurements. For one, this number is only a probabilistic data point, referring to how likely it is that the particular treatment group is better off than the control group. For another, the number is

by no means a hard fact from which a clear policy solution automatically follows. Instead, the number needs to be interpreted and applied to the policy issue in question.

(b) Understand the qualitative-interpretive components of RCTs

Pairing RCT evaluations with interpretive components is not only a requirement, but part of its design and application. Every RCT relies heavily on qualitative-interpretive components throughout the evaluation process. This includes analyzing whatever prior research is available in the policy field, which requires a judgment of which studies are relevant and evaluating those studies for their relevance. Such background research is essential in order to clarify the policy issue under study and to plan the evaluation design. RCT theorist Ronald Fisher argued that every RCT needed to start with the selection of an explicitly formulated hypothesis, which was always inductive and thus subject to qualitative-interpretive reasoning (Fisher, 1933, 9). Qualitative skills assist in deciding on the sample size, which in turn depends on the sample heterogeneity and the expected size of the effect—both of these are estimated rather than known. Qualitative skills are required to predict potential causal effects, to determine which baseline measures are relevant, and to understand the political and social environment where the experiment is to be performed. Furthermore, after running the statistical-experimental model, the evaluator needs to interpret RCT findings and generate policy conclusions. These skills contain qualitative elements that cannot be derived from pure cause-effect quantification. Interpretive reasoning helps make RCT findings relevant for policy makers; it can set the boundaries in which the findings are most likely applicable and determine how representative those findings are—an issue separately covered in the next recommendation. In fact, the quality of an RCT directly depends on the appropriate utilization of interpretive-qualitative skills to determine program impact.

Frequently, RCT evaluators are unaware that they are not merely required to perform a quantitative evaluation, but that they must also use qualitative-interpretive skills when planning and executing an RCT. They need to become more aware of these interpretive-qualitative components of their evaluations so that they can be positively utilized. Policy makers also need to recognize that every RCT finding is based on qualitative-interpretive reasoning.

(c) Clarify representativeness of RCT findings

Program evaluation as an applied science is, per definition, interested in producing research relevant to the larger society, thereby applying findings to contexts beyond the studied population. Application of RCT findings to other policy contexts requires establishing a qualitative theory of equivalency. This means determining the relations that need to hold between the RCT sample population and the population in the new policy context. This could also be called the transferability or translation of RCT findings to other policy contexts. Fisher emphasized that no isolated experiment, however significant in itself, could suffice for demonstrating the general effectiveness of an intervention (Fisher, 1935, 16). The medical statistician Hill argued that to extrapolate from a sample to the general run of patients, one would need to carefully consider if the sample was fully representative of all patients, and not in any way biased or selected (Hill, 1937, 9).

I find the most lucid statement of the problem argued by the philosopher of science, Nancy Cartwright. She analyzes this problem of equivalency in her work on external validity (Cartwright & Munro, 2010; Cartwright, 2009, Cartwright, 2007): although RCTs may be internally valid, they alone do not generate any external validity. Additional interpretive-qualitative reasoning is necessary to generate externally valid findings. However, many RCT evaluators stop with the determination of internal validity, or make statements about general representativeness of findings without basing these claims on systematic reasoning. In contrast, policy makers need to carefully test RCT findings to determine the degree to which they are applicable to the new policy context.

A related issue is moving from a pilot experiment to a large implementation of a program. As shown in chapter two, the Poliomyelitis experiments were subjected to stringent oversight in the vaccine production, whereas the nation-wide implementation a year later used a less stringent process, due to the exponential increase in demand for that vaccine. As a result, the vaccines used in the Salk trial and the vaccines used in the general distribution differed and resulted in higher infection rates in the latter group. This problem of re-applying pilot findings is even larger in educational systems with more factors to consider. The Tennessee Class Size Reduction experiment, for example,

provided statistically significant findings for the sample population of seventy-nine schools narrowed down by size and facility, but not necessarily for a more general population. The challenge is determining in what other contexts one would obtain equivalent, or at least similar, findings. Put another way, what components of the original tested intervention and its policy context would be required to guarantee the applicability of findings in a new policy context?

Heterogeneous populations and intervention components under study may increase representativeness. When designing an RCT, an important question is: How much heterogeneity does the evaluator allow without jeopardizing statistically significant findings? In the discipline of community health, Penelope Hawe and coauthors have taken a provocative stance in arguing for what they call "out-of-control" trials—so-called because they are not restricted to a homogeneous population and program structure (Hawe, Shiell, & Riley, 2004). They argue that RCTs would be suitable for complex interventions, when one allows for heterogeneity in, for example, intervention characteristics and treatment units, *inter alia*. For instance, one single RCT could include various obesity programs, such as personal counseling or group counseling. These counseling services might be offered to various clients, such as middle-aged males or teenage girls. Hawe et al.'s argument countered RCT skeptics (e.g., NONIE's Alternative Group), who opposed RCTs for complex interventions. These skeptics argued that RCTs were only appropriate for single-strand initiatives, with a concrete intervention and explicit expected outcomes, such as body height. Hawe et al. argued, however, that RCTs could still be performed in heterogeneous treatment and control groups. In fact, they would create a more realistic picture of real-world situations. However, I would argue that "out-of-control" RCTs have major drawbacks. Heterogeneity in the implementation of an intervention, specifically of the subjects and their context, could lead to greater uncertainty as to why an intervention worked. Furthermore, less control could mean a lower chance of significant findings due to the heterogeneity of subjects, contexts, and interventions. The evaluator thus needs to find a compromise between heterogeneous and homogeneous samples, interventions, and contexts to account for representativeness of RCT findings.

(d) Expand the evidence base beyond the RCT

In the 1980s, Lee Cronbach had already challenged the preference given to RCT approaches to guarantee internal validity, and he emphasized the need for generalizability of results. He therefore argued for correlational approaches, whose predictions are better tailored to "real life" and offer better generalizability (Wittmann, Klumb, 2006).

Evaluators may also look beyond experimental approaches to determine program impact. Impact evaluations rely on a simulated counterfactual, either explicit or implicit. Experimental evaluations as well as qualitative-interpretive evaluations and quantitative single-case studies are predicated on the notion of a counterfactual. A counterfactual concept is important because it deals with possible exogenous factors that may affect the program outcomes. For example, a sick person might be showing relief of symptoms via the bodily healing process, even when not receiving a certain drug. The Streptomycin trial found some cases where patients improved their health without the antibiotics. Campbell and Stanley called this effect "maturation" (Campbell & Stanley, 1967). Repeated single-case studies at various locations and times reduce the threat of such maturation by subjecting a particular unit to an intervention at different points in time. They thus approximate the counterfactual diachronically. Since 2010, the What Works Clearinghouse recognizes single-case studies as a valid approach in determining program impact (Kratochwill et al., June 2010).

The fields of education and international development would also be well advised to learn from medicine. Large-scale RCTs cover only a small portion of medical research. This small proportion is because thousands of chemical compounds exist, but only a few can be tested by RCTs. Pre-clinical tests need to make a first cut in separating the promising from the less promising compounds. Qualitative-interpretive observations may serve a similar purpose in other policy fields, by allowing the selection of promising interventions in education and international development.

Furthermore, personalized medicine and translational approaches have been on the rise, moving from the bench (i.e., the laboratory) to the bedside, where systematic observational evidence in collaboration with medical practitioners is used. Observational

data from preclinical, phase 1, and phase 2 trials contribute to determining causal effects of medications (Solomon, presentation, June 19, 2009). These trials often use a general elimination approach to systematically test the competing explanations of program effects (Scriven, 2008). These and other designs are important complements to the phase 3 textbook RCT. They broaden the evidence base, and thus the scientific base, for policy decisions.

(e) Produce evaluations on comparative effectiveness

Comparative effectiveness research (CER) in health could serve as a frontrunner approach for education and international development. CER is a recent movement that emphasizes the comparability of findings. Rather than using non-treatment control groups, CER requires that the control group receive the best available treatment on the market. For example, any new anti-depressive medication would need to be compared to the best currently available antidepressant. Similarly, in education and international development, RCTs could compare novel interventions to the best available known treatment. For example, rather than comparing reduced class-size to regular class size, the evaluators could compare class-size reduction to an extended-school-day program. Or, schools districts could compare in-classroom reading programs with small group tutoring programs. These types of comparisons would be more meaningful for policy makers, who must choose between competing interventions within budgetary constraints. Not only would these results be more relevant for decision makers, but they would also be more ethical because both groups could potentially benefit from the program.

CER also emphasizes the generalizability of findings to the general run of patients, as Hill would have put it. CER relies on pragmatic trials, i.e., trials in routine clinical practice, and also it values observational findings and modeling. CER thus provides greater methodological flexibility and does not necessarily privilege the RCT over all other methods.

CER is also concerned with heterogeneous patient characteristics and subgroup distinctions. RCT researchers, however, do not use subgroup analysis when the original random assignment has not taken place according to these subgroups. In contrast, CER

attempts to include subgroup analysis and thereby strengthens externally valid results. In education and international development, CER could encourage subgroup analysis, while statistically accounting for the change in the unit of analysis. CER would help identify differences of findings among beneficiary groups and thus would contribute to more relevant findings for contextual policymaking.

The cost factor can also be part of comparative effectiveness research. The question then goes beyond whether "the intervention [is] effective," and asks if it is worth the money invested (Wittmann, 2011). This answer is evaluative-judgmental, going beyond the RCT methodology, and incorporating qualitative reasoning.

(f) Strengthen the evaluation function

Evaluation findings should assume a more relevant role in the policy-making process. One condition for doing so would be to increase the credibility of the evaluation function. Policy makers could be trained in how to utilize evaluation findings so that they could gain a realistic perspective on the strengths and limitations of evaluations and RCT evaluations in particular. In turn, evaluators could prominently include caveats on RCT findings in evaluation reports, especially regarding their representativeness for other policy contexts. Policy makers might then be less apt to use findings inappropriately.

Regarding the evaluation process itself, evaluation teams should reflect interdisciplinary and methodological diversity to provide well-balanced impact evaluations. The evaluation teams would ideally include methodological specialists for both experimental-quantitative and observational-qualitative approaches. Such integration would lead to a more balanced evaluation product.

Progresa is a successful example how well-trained scholars transformed into influential practitioners who played a fundamental role in promoting the new conceptual approach of poverty reduction (Lustig, 2011, 2). These scholar-practitioners ensured the technical soundness and effectiveness of the program's design; they incorporated rigorous impact evaluations in the program's design; and they ultimately persuaded politicians to implement and keep the program in place (cf., chapter four). Despite all its limitations in

evidence, Conditional Cash Transfer became an important policy tool in poverty reduction. The evaluations were RCTs supplemented with strong qualitative approaches. Progresa is an exemplary story for the role of high-quality evaluations in policy-making.

Synthetics reviews and meta-analyses of experimental and non-experimental findings should become more integrated in the policy review process. Organizations such as the Campbell Collaboration, or the International Initiative for Impact Evaluations, are prime examples of providing such syntheses on scientific findings in certain policy domains. Scientific reviews of evaluation findings could help in the utilization of impact evaluations in the political decision-making process. Ultimately, impact evaluations are only as valuable to the extent that they are able to inform the policy process. Based on these recommendations, impact evaluations could obtain the necessary level of influence on policy decisions.

This study contributes to strengthening the evaluation function by providing a realistic, analytical assessment of the RCT as an evaluation tool to determine program impact in three distinct policy areas. The study's principal contribution is an interpretive review of the RCT use and debates and how the RCT evolved and adapted from medicine, on the one hand, to education and international development, on the other. In particular, the study investigates the notion of the RCT as a seamless success story in medicine and how this notion underpins the use of RCTs in transforming other policy areas such as education and international development. Whenever RCTs become part of a political debate, policy makers and decision makers are reminded that even RCT results can be biased and political, despite their scientific, unbiased aura.

5. Limitations of the study

As is generally accepted, scientific knowledge is provisional and impermanent. Theories of cause and effect need to be updated and continuously based on new evidence, but by definition they can never be proven to be true. Equally, generalizations are necessarily uncertain, and it is often unclear how to determine the extent to which research findings may be generalized between fields. These limitations also apply to my research study.

It was necessary to make spatial-temporal interpolations, because I traced a methodological tool across time and policy domains. The documents I used were always specific data points situated in a set time and set space. These cases were only a few of many in which a concept such as the RCT manifests itself, and I had to make interpolations from one spatial-temporal context to another. The Center for Global Development's (CGD) report, for instance, was published in May 2006. In December 2007, the European Evaluation Society (EES) released a statement about the inappropriate use of experimental designs in impact evaluation. Although EES did not directly refer to the CGD study, I was able to link these documents; I had spoken to individuals who had participated in the drafting of the EES statement, in which they clearly made reference to the CGD study. I induced evidence from a limited number of contexts to make my case that the EES statement was a direct response to the CGD report. I chose a certain narrative and selected available sources as evidence to make my case, but other interpretations may be possible when using additional data. My interpretation is provisional and may change, based on new evidence. This possibility of changing findings is the nature of a hermeneutic-interpretive approach, just as new findings interact with a quantitative or experimental approach.

My research study is limited to the RCT approach in three policy disciplines: medicine, education, and international development, and it should be read within these realms. The RCT movement, however, has spread to other disciplines, where it has been absorbed and discussed along different lines. For example, criminal justice has also relied on RCT evidence since the 1990s and has not faced much criticism or backlash (Sherman, 1998, 4; Sherman, 2002). On the contrary, in education and international development, many evaluators criticized the RCT approach, as I showed in chapters three and four. Therefore, findings about the RCT movements may be different in other policy fields.

Furthermore, this study focuses more on the methodological and practical challenges rather than on the political realities of using the RCT approach. The analysis of the Reading First Initiative and its state-level responses hint at the political consequences of implementing a so-called scientifically based research agenda, but they are not fully developed in this study. I opted for breadth rather than depth to accommodate a cross-

disciplinary perspective. However, the rhetoric associated with the medical RCT as a success story reveals a desire for science to serve as an apolitical tool by enabling objective and unbiased policy making. As the example of Reading First illustrated, this adopting of science became a political tool legitimizing certain (phonics-based) practices in schools and school districts. Future research would benefit from further analysis of the political dimensions associated with using RCTs and quantitative methods more generally.

6. Future Research

Future research should comparatively explore these RCT movements across different cultures and in additional policy areas. For the field of medicine in the 1970s, Archibald Cochrane argued that the degree of RCT use had a geographical distribution: There was a high use of RCTs in the United States, United Kingdom, and Scandinavia, but almost none in Catholic, Communist, or underdeveloped countries (Cochrane, 1971). Although Cochrane's claim is now 40 years old, it would be worthwhile to explore a modernized version of his view—asking in effect what the cultural, socio-political, and institutional factors are that lead to the promotion of RCTs in evaluation. Cochrane mentioned "authoritarianism" as a possible cause for a lack of RCTs. Cochrane's question should be further explored.

Although this study focuses on the RCT approach, an expansion of methodological approaches is necessary to effectively evaluate program impact. Many macroeconomic and macro-policy interventions do not allow for a random assignment of individual units, but rather deal with one unit, such as a national economy or an institutional change. Although a policy intervention might affect individuals and households, its primary influence is on institutions and processes. A control-group construction is not feasible as macroeconomic policies affect everyone. Therefore, it is necessary to develop and apply non-experimental methods to these macro-level interventions. Future research needs to focus on how non-experimental and non-quantitative methods could be used to determine program impact.

APPENDIX

1. Abbreviations

3IE: International Initiative for Impact Evaluation

AEA: American Evaluation Association

CER: Comparative Effectiveness Research

CGD: Center for Global Development

ECG: Evaluation Cooperation Group of the multilateral financial institutions

EES: European Evaluation Society

ESEA: Elementary and Secondary Education Act

EBM: Evidence-Based Medicine

FDA: United States Food and Drug Administration

GAO: United States Government Accountability Office

ICC: International Cochrane Collaboration

ICCE: International and Cross-Cultural Evaluation (Topical Interest Group of the American Evaluation Association)

ITS: Interrupted time series

IOCE: International Organization for Collaboration in Evaluation

LFA: Logical Framework Approach

NCLB: No Child Left Behind Act

NDA: New Drug Application

NICE: United Kingdom's National Institute for Health and Clinical Excellence

NRP: National Reading Panel

NONIE: Network of Network on Impact Evaluation

MRC: United Kingdom's Medical Research Council

NAEP: National Assessment for Educational Progress

OECD DAC: Organization for Economic Cooperation and Development's Development Assistance Committee

PRA: Participatory Rural Appraisal

RCT: Randomized Controlled Trial

STAR: Student Teacher Achievement Ratio (=Tennessee class size reduction trial)

TED: Technology, Education, and Development series

UNEG: United Nations Evaluation Group

USAID: United States Agency of International Development

USDOE: United States Department of Education

WWC: What Works Clearinghouse of the USDOE

2. Definitions

Definitions are based on the OECD Development Assistance Committee's *Glossary of key terms in evaluation and results based management* (2002) and the Research Methods Knowledge Base by William M.K. Trochim (www.socialresearchmethods.net).

Activity: Actions taken or work performed through which inputs, such as funds, technical assistance and other types of resources, are mobilized to produce specific outputs.

Attribution: The ascription of a causal link between changes and a specific intervention.

Attrition: Loss of subjects from the defined sample during the course of a study.

Bias: A systematic difference from the population parameter of interest.

Control group: A randomly assigned group as closely as possible equivalent to an experimental group (one that is exposed to a program), and exposed to all the conditions of the investigation except the program being studied.

Counterfactual: The situation which hypothetically may prevail for individuals, organizations, or groups were there no intervention.

Effect: Change due to an intervention.

Effectiveness: The extent to which the intervention's objectives were achieved.

Efficiency: A measure of how economically resources/inputs (funds, expertise, time, etc.) are converted to results.

External validity (cf., generalizability): The degree to which the conclusions in a study would hold for other persons in other places and at other times.

Evaluation: The systematic and objective assessment of an on-going or completed project, program or policy, its design, implementation and results. The aim is to determine the relevance and fulfillment of objectives, development efficiency, effectiveness, impact and sustainability. An evaluation should provide information that is credible and useful, enabling the incorporation of lessons learned into the decision-making process of both recipients and donors.

Experiment (narrower sense): Randomized controlled trial.

Experiment (broader sense): Scientific investigation in which an investigator manipulates and controls one or more independent variables to determine their effects on the outcome (dependent) variable.

Generalizability (cf., external validity): The extent to which information about a program collected in one setting can be used to reach a valid judgment about how it will perform in other settings.

Impact: Positive and negative, primary and secondary long-term effects produced by a development intervention, directly or indirectly, intended or unintended.

Input: The financial, human, and material resources used for the development intervention.

Internal validity: The approximate truth about inferences regarding cause-effect or causal relationships.

Level of significance: The probability that the observed difference occurred by chance.

Logic Model: Displays the sequence of actions that describe what the program is and will do: how inputs link to activities, outputs, outcomes, and impact.

Logical Framework Approach: Management tool used to improve the design of interventions, most often at the project level. It involves identifying strategic elements (inputs, outputs, outcomes, impact) and their causal relationships, indicators, and the assumptions or risks that may influence success and failure. It thus facilitates planning, execution, and evaluation of a development intervention.

Output: The products, capital goods, and services which result from a development intervention; may also include changes resulting from the intervention which are relevant to the achievement of outcomes.

Participatory evaluation: Evaluation method in which representatives of agencies and stakeholders (including beneficiaries) work together in designing, carrying out, and interpreting an evaluation.

Qualitative data: Facts and claims presented in narrative, not numerical, form.

Quantitative data: Facts and claims that are represented by numbers.

Quasi-experiment: A quasi-experimental design is one that looks a bit like an experimental design but lacks the key ingredient, i.e., random assignment.

Randomized controlled trial (RCT): Study in which units (i.e., people, households, communities) are allocated at random to receive one of several interventions.

Regression discontinuity design: Participants are assigned to treatment or comparison groups on the basis of a cutoff score on a pre-program measure.

Results chain: The causal sequence for an intervention that stipulates the necessary sequence to achieve desired objectives beginning with inputs, moving through activities and outputs, and culminating in outcomes, impacts, and feedback.

Selection bias: Any factor other than the program that leads to posttest differences between groups.

Systematic review: The purpose is to sum up the best available research on a specific question by synthesizing the results of several studies (Campbell Collaboration).

Triangulation: The use of multiple sources and methods to gather similar information.

Unit of analysis: The least divisible element on which measures are taken and analyzed.

Validity: The soundness of the use and interpretation of a measure.

BIBLIOGRAPHY

The citations are organized by text format, as follows:

1. Postings on electronic mailing lists (list serves)
2. Electronic blogs & press releases
3. Speeches, conference presentations
4. Legislation and legislative hearings
5. Meeting notes
6. Interviews, electronic communications
7. Newspapers, magazines
8. Reports
9. Journal articles
10. Monographs and book chapters

1. Postings on electronic mailing lists (list serves)

Burke, A. (May 30, 2008). Consequences of No Child Left Behind for educational evaluation. Message posted to Evaltalk, archived at bama.ua.edu/archives/evaltalk.html.

Krueger, R. (November 24, 2003). AEA response to scientific based evaluation methods. Message posted to Evaltalk, archived at bama.ua.edu/archives/evaltalk.html.

Lipsey, M. W. (December 3, 2003). NOT the AEA statement on scientifically based evaluation. Message posted to Evaltalk, archived at bama.ua.edu/archives/evaltalk.html.

Lipsey, M. W. (December 18, 2003). Methodological pluralism. Message posted to Evaltalk, archived at bama.ua.edu/archives/evaltalk.html.

Lipsey, M. W. (December 11, 2003). AEA statement on research methods. Message posted to Evaltalk, archived at bama.ua.edu/archives/evaltalk.html.

Patton, M. Q. (September 1, 2003). Experiments. Message posted to Evaltalk, archived at bama.ua.edu/archives/evaltalk.html.

Perrin, B. (December 11, 2003). AEA statement on research methods. Message posted to Evaltalk, archived at bama.ua.edu/archives/evaltalk.html.

Scriven, M. (November 12, 2003). Comments to Feds re proof of causation. Message posted to Evaltalk, archived at bama.ua.edu/archives/evaltalk.html.

Scriven, M. (May 1, 2011). USAID's new evaluation policy. Message posted to Evaltalk, archived at bama.ua.edu/archives/evaltalk.html.

Williams, B. (Jan 9, 2007). Evaluation of social programs worldwide. Message posted to Evaltalk, archived at bama.ua.edu/archives/evaltalk.html.

Wittmann, W. W. (December 5, 2003). Re: NOT the AEA statement on scientifically based evaluation. Message posted to Evaltalk, archived at bama.ua.edu/archives/evaltalk.html.

2. Electronic blogs & press releases

American Educational Research Association (December 3, 2003). *Resolution on the essential elements of scientifically-based research.* Retrieved from http://www.eval.org/doeaera.htm.

American Evaluation Association (December 3, 2003). *Evaluation leaders decry Department of Education's proposed evaluation methods. Press release.* Retrieved from www.eval.org/doe.pressrelease.htm.

American Evaluation Association (November 25, 2003). *Response to the U. S. Department of Education notice of proposed priority "Scientifically Based Evaluation Methods".* Retrieved from http://www.eval.org/doestatement.htm.

Barder, O. (September 5, 2011). Evolution and complexity in development evaluation, Retrieved from http://media.owen.org/Evolution/player.html.

Center for Global Development (June 12, 2006). *A major step forward on impact evaluation. Press release.*

Center for Global Development (June 8, 2006). *Donors back impact evaluation of programs in developing countries: New initiative to support independent studies to determine what works.* Bellagio, Italy.

Charlotte Chamber of Commerce (2001). *CMS partners for school reform. What works in reading?* Retrieved from www.charlottechamber.com.

Cuthbertson, L. (2008). *Royal College of Physicians: Sir Michael Rawlins attacks traditional ways of assessing evidence: Press release,* Retrieved from http://www.politics.co.uk.

European Evaluation Society (December 2007). *EES Statement: The importance of a methodologically diverse approach to impact evaluation. Specifically with respect to development aid and development interventions.* Retrieved from http://www.ees.kingsquare.nl/download/?noGzip=1&id=1969403.

Institute of Education Sciences (May 17, 2003). Whatworksclearinghouse.org. Retrieved from www.archive.org.

National Education Association (December 3, 2003). *Comments on scientifically based evaluation methods.*

Shanahan, T. (1999). *The National Reading Panel: Using research to create more literate students: An invited contribution.* Retrieved from http://www.readingonline.org.

Trochim, W. M. K. (2006). *Research Methods Knowledge Base.* Retrieved from http://www.socialresearchmethods.net/kb/.

U.S. Department of Education (August 2, 2002). *Frequently asked questions and answers for families and communities.* Retrieved from www.archive.org.

3. Speeches, conference presentations

Adato, M. (November 24, 2008). *Using mixed methods to improve evaluation of conditional cash transfer programs in Mexico, Nicaragua and Turkey.* Belo Horizente, Brazil.

Barder, O. (July 2011). *Can aid work? Written testimony submitted to the House of Lords.* Washington, DC: Center for Global Development.

Baron, J. (November 8, 2008). In Rigorous evidence: The key to progress in education? Lessons from medicine, welfare and other fields. Forum Proceedings, Ed. Coalition for Evidence-Based Policy, Washington, DC.

Bush, G. W. (January 8, 2002). *President Signs Landmark No Child Left Behind Education Bill.* Retrieved from http://georgewbush-whitehouse.archives.gov/news/releases/2002/01/20020108-1.html.

Clinton, W. J. (February 4, 1997). *1997 Address Before a Joint Session of the Congress on the State of the Union*. Retrieved from http://www.presidency.ucsb.edu/medialist.php?presid=42.

Duflo, E. (February 1, 2010). *Social experiments to fight poverty.* Technology, Education, Design (TED) Talk, Long Beach, CA. Retrieved from http://www.ted.com/talks/esther_duflo_social_experiments_to_fight_poverty.html.

Easton, J. Q. (March 28, 2011). *How relevant research can help build a science of education.* University of North Carolina at Chapel Hill and Frank Porter Graham Institute Talk. Chapel Hill, NC.

Fisher, R. A. (1931). *Principles of plot experimentation in relation to the statistical interpretation of the results.* Rothamsted, from Conference on The Technique of Field Experiments.

Fiszbein, A. (January 16, 2008). *Learning from development practice to improve policy and program design.* Washington, DC.

Gaarder, M. (May 4, 2010). *Conditional cash transfers and health: Unpacking the causal chain.* Washington DC: Center for Global Development. Retrieved from http://www.3ieimpact.org/userfiles/doc/Gaarder_II_May_4th_CCT_and_health_clean[1]%20[Compatibility%20Mode].pdf.

Gertler, J. (January 13, 2008). *Measuring results and impact evaluation. From promises into evidence.* SIEF Impact Evaluation Workshops: Cairo, Egypt. Retrieved from http://go.worldbank.org/AE8QGH8HX0.

Glennerster, R. (November 10, 2008). *Fighting poverty: What works? Running randomized evaluations of poverty programs in developing countries.* Baltimore, MD: American Evaluation Association.

Greene, J. C. (April 2, 2009). Design alternatives for impact evaluation and for generating "credible evidence." Cairo, Egypt: Perspectives on Impact Evaluation Conference.

Hamburg, M. A. (September 27, 2010). *Better health through biomedicine: Innovative governance. Workshop in Berlin, Germany*. Retrieved from http://www.fda.gov/NewsEvents/Speeches/ucm242043.htm.

Hamburg, M. A. (May 15, 2010). *Innovation and the FDA: Past, present, and future. Massachusetts Medical Society's 2010 Shattuck Lecture.* Retrieved from http://www.fda.gov/NewsEvents/Speeches/ucm242008.htm.

Hamburg, M. A. (February 25, 2010). *Remarks at Personalized Medicine Coalition's Sixth Annual Keynote Luncheon.* Retrieved from http://www.fda.gov/NewsEvents/Speeches/ucm210018.htm.

Hamburg, M. A. (October 13, 2010). *Remarks at the Academy of Medical Sciences, England,* London.

Heinrich, C. (April 18, 2009). Comparative overview of conditional cash transfers, Presented at The Origins, Implementation, and Spread of Conditional Cash Transfer Programs in Latin America, The University of Texas at Austin: Austin, TX.

Levine, R. (April 28, 2010). *Discussion points.* Presented at Impact Evaluation Clinic*,* Washington, DC: InterAction.

Lewis, M. (January 16, 2008). *Spanish Trust Fund for Impact Evaluation (SIEF).* Presented at Making Smart Policy: Using Impact Evaluation for Policy Making, Washington, DC: World Bank.

Martinez, S., & Gertler, J. (January 14, 2008). *Measuring impact: Impact evaluation methods for policy makers.* Cairo, Egypt: World Bank.

NONIE (May 24–25, 2007). *NONIE Management: A proposal.* The Hague, Netherlands: Network of Networks on Impact Evaluation.

NONIE Subgroup 1 (May 24–25, 2007). *Experimental and quasi-experimental approaches to impact evaluation: Moving forward to action.* The Hague, Netherlands: Network of Networks on Impact Evaluation.

NONIE Subgroup 2 (May 24–25, 2007). *Approaches and methods in impact evaluation.* The Hague, Netherlands: Network of Networks on Impact Evaluation.

NONIE Subgroup 2 (January 14, 2008). *NONIE impact evaluation guidance. Presentation.* Washington, DC: Network of Network on Impact Evaluation.

Peterson, P. (December 8, 1999). *What constitutes high quality education research and how can it be incorporated into policymaking? Can we make education policy on the basis of evidence?* Transcript. Washington, DC: Brookings Forum on Education Policy.

Solomon, M. (June 19, 2009). *Just a paradigm: Evidence-based medicine meets philosophy of science.* Paper presented at the Society for Philosophy of Science in Practice, Minneapolis MN.

Warner, A. (October 3, 2008). *Overcoming intellectual sub-cultures in Development Evaluation – the experience of NONIE.* Lisbon, Portugal: European Evaluation Society.

Vinod, T. (February 20, 2008). *Impact evaluation: Its status and its future: 7th meeting.* Paris, France: OECD Development Assistance Committee Network on Development Evaluation.

White, H. (April 28, 2010). *Remarks.* Presented at Impact Evaluation Clinic*,* Washington, DC: InterAction.

Zedillo, E. (April 17, 2009). *A look at Progresa's genesis.* Presented at The Origins, Implementation, and Spread of Conditional Cash Transfer Programs in Latin America, The University of Texas at Austin: Austin, TX.

4. Legislation and legislative hearings

89th United States Congress (April 11, 1965). Elementary and Secondary Education Act. Public Law 89-10, 79 Stat. 27.

96th United States Congress (October 17, 1979). Department of Education Organization Act. Public Law 96-88, 93 Stat. 668.

107th United States Congress (January 8, 2002). No Child Left Behind Act of 2001. Public Law 107-110, 115 Stat. 1425.

107th United States Congress (November 5, 2002). Education Science Reform Act. Public Law 107-279, 116 Stat. 1940.

105th United States Congress (October 21, 1998) Reading Excellence Act. A part of the Omnibus Consolidated and Emergency Supplemental Appropriations Act. Public Law 105-277, 112 Stat. 2681.

Congressional Record House (November 7, 2007). *Conference report on H.R. 2264, Department of Labor, Health and Human Services, and Education, and related agencies appropriations* H10230.

Food and Drug Administration (1970). Hearing regulations and regulations describing scientific content of adequate and well-controlled clinical investigations. *Federal Register, 35.90: 7250-7253*.

Alexander, D. (June 19, 1997). *Testimony*. *A bill to establish a National Panel on Early Reading Research and Effective Reading Instruction,* to the Committee on Labor and Human Resources. Congressional Record Senate, S6003. Retrieved from www.gpo.gov.

An Act to amend the Tennessee Code Annotated (1985). Title 49, Chapter 3, relative to incentives for class size reductions, Section 1.

Langenberg, D. N. (April 13, 2000). *Statement. Report of the National Reading Panel. Hearing before a subcommittee of the committee on appropriations.* 106th Congress 2nd session. Special hearing. Retrieved from www.gpo.gov.

Lyon, G. R. (October 26, 1999). *Testimony. Education research. Is what we don't know hurting our children? Hearing before the House Committee on Science, Subcommittee on Basic Research.* Retrieved from www.gpo.gov.

Lyon, G. R. (March 8, 2001). *Testimony. Hearing on measuring success: Using assessments and accountability to raise student achievement before the House Committee on Education and the Workforce, Subcommittee on Education Reform.* Retrieved from www.gpo.gov.

OECD (March 2, 2005). *Paris Declaration on aid effectiveness.* Ownership, harmonisation, alignment, results and mutual accountability. Paris, France: Organization of Economic Cooperation and Development.

U.S. Department of Education (2003). *Notice of proposed priority. Scientifically based evaluation methods:* Federal Register 68.213.

U.S. Department of Education (January 25, 2005). *Notice. Scientifically based evaluation methods:* Federal Register: 70.15.

United Nations (September 8, 2000). *United Nations Millennium Declaration.*

Whitehurst, Grover (Russ) (June 25, 2002). *Testimony. Hearing of the Committee on Health, Education, Labor, and Pensions.* United States Senate. Examining proposed legislation authorizing funds for the Office of Education Research and Improvement (OERI). Retrieved from www.gpo.gov.

5. Meeting notes

National Reading Panel (April 24, 1998). *Inaugural meeting. Meeting minutes.* Bethesda, MD. Retrieved from http://www.nationalreadingpanel.org/nrpabout/Meetings_Archive.htm.

National Reading Panel (September 10, 1998).*Meeting minutes.* Washington, DC. Retrieved from http://www.nationalreadingpanel.org/nrpabout/Meetings_Archive.htm.

National Reading Panel (October 19, 1998). *Subgroup Chairs Meeting. Meeting minutes.* Bethesda, MD. Retrieved from http://www.nationalreadingpanel.org/nrpabout/Meetings_Archive.htm.

National Reading Panel (November 9–10, 1998). *Meeting minutes.* Washington, DC. Retrieved from http://www.nationalreadingpanel.org/nrpabout/Meetings_Archive.htm.

NONIE notes (January 14, 2008). *Meeting notes,* taken by Rahel Kahlert. Washington, DC.

Network of Networks on Impact Evaluation Steering Committee (September 9, 2008). *Summary notes.*

Network of Networks on Impact Evaluation (November 15, 2006). *Impact evaluation workshop. Summary notes.* Paris.

6. Interviews, electronic communications

Crepon, B. (June 18, 2008). Challenges of RCTs, Jameel Poverty Action Lab Europe, Paris, France.

Bryk, Anthony (January 2, 2011). Lehrer–Scientific Method 2010, email communication with Uri Treisman.

Guijt, I. (April 2010). Reflections on the Cairo Impact Evaluation Conference. Phone conversation.

Lipsey, M. W. (October 3, 2011). Response to USDOE priority, email communication.

Ofir, Z. (August 6, 2007). On RCT movements in education and international development, email communication.

Patton, M. Q. (November 3, 2006). Experimental evaluations. American Evaluation Association, Portland, OR.

Rogers, P. (April 29, 2010). On NONIE guidance, InterAction Impact Evaluation Workshop, Washington, DC.

Rugh, J. (November 3, 2007). On the CGD Report on Impact Evaluation. American Evaluation Association, Baltimore, MD.

Russon, C. (April 8, 2008). On the Evaluation Gap Working Group. Email conversation.

Savedoff, W. (January 17, 2008). On the Evaluation Gap Working Group, Washington, DC.

White, H. (April 28, 2010). On the Evaluation Gap Working Group. InterAction Impact Evaluation Workshop. Washington, DC.

7. Newspapers, magazines

Banerjee, A. V. (July/August 2006). *Making aid work. Boston review.* Retrieved from http://bostonreview.net/BR31.4/banerjee.php.

Dugger, C. (July 28, 2004). World Bank challenged: Are the poor really helped? *New York Times*.

Education Week (September 8, 2004). Select group ushers in reading policy; 'Evidence based' drive led by head of federal child-research branch. *Education Week,* 2.

Education Week (September 7, 2005). States pressed to refashion Reading First grant designs. Documents suggest federal interference. *Education Week*, 1.

Grunwald, M. (October 1, 2006). Billions for an inside game on reading. *Washington Post*, B1.

Krueger, A. B. (May 2, 2002). Putting development dollars in use, south of the border. *New York Times*.

Manzo, K. K. (December 12, 2006). Reading law fails to bring innovations. *Education Week, 26*(15), 1–13.

Manzo, K. K. (March 6, 2007). Ed. Dept. allowed singling out of 'Reading First' products. *Education Week, 26*(26), 13.

MIT Technology Review (January 2010). *Poverty's researcher.*

Olson, L., & Viadero, D. (January 30, 2002). Law mandates scientific base for research. *Education Week, 21*(20), 14–15.

Orlich, D. C. (1991). Brown v. board of education: Time for reassessment. *Phi Delta Kappan, 72*(8), 631–632.

Pomeranz, D. (May 13, 2008). Wirtschaftswissenschaft im Dienste der Armen. Ökonomen propagieren neue Evaluationsmethoden in der Entwicklungshilfe. *Neue Zürcher Zeitung*.

Ravallion, M. (February 2009). Should the randomistas rule?: Economists' voice. *The Berkeley Electronic Press.*

Rodrik, D. (June 12, 2008). Control freaks. Are "randomized evaluations" a better way of doing aid and development policy? *The Economist*.

Strom, S. (July 13, 2003). Gates aims billions to attack illnesses of the world's neediest. *New York Times*, 1.

Strom, S. (July 11, 2008). Some philanthropists are no longer content to work quietly. *New York Times*, 21.

Try and see it (February 28, 2002). *The Economist*, 73.

Zigler, E. (1992, June 27). Head Start falls behind. *New York Times*, 15.

8. Reports

Achilles, C. M. (January 14, 1993). *The Lasting Benefits Study (LBS) in grades 4 and 5 (1990-1991): A legacy from Tennessee's four-year (K–3) class size study (1985-1989), project STAR.* Greensboro, NC.

Bamberger, M. (2006). *Conducting quality impact evaluations under budget, time and data constraints.* Washington, DC: The World Bank.

Benhassine, N., Devoto, F., Duflo, E., Dupas, P., & Pouliquen, V. (2010). *The impact of conditional cash transfers on schooling and learning: Preliminary evidence from Tayssir pilot in Morocco.*

Behrman, J. R. (2007). *Policy-oriented research impact assessment case study on the International Food Policy Research Institute and the Mexican Progresa anti-poverty and human resource investment conditional cash transfer program*, Impact Assessment Discussion Paper 27. Washington, DC: International Food Policy Research Institute.

Behrman, J. R., Parker, S., & Todd, D. (2005). *Long-term impacts of the Oportunidades Conditional Cash Transfer Program on rural youth in Mexico*, Discussion Paper 122. Göttingen, Germany: Ibero-America Institute for Economic Research.

Bohrnstedt, G. W., & Stecher, B. M. (2002). *The Capstone Report: What we have learned about class size reduction in California.* Sacramento, CA: California Department of Education.

Bourguignon, F., & Sundberg, M. (2007). *Aid effectiveness. Opening the black box.* United Kingdom: Governance and Social Development Resource Centre.

Duncan, R. (2008). *Evaluation in developing countries: What have we learned?* Unpublished paper.

Evaluation Gap Working Group (2006). *When will we ever learn? Improving lives through impact evaluation.* Washington, DC: Center for Global Development.

Savedoff, W., Levine, R., & Birdsall, N. (September 15, 2005). *When will we ever learn? Recommendations to improve social development through enhanced impact evaluation. Consultation draft.* Washington, DC: Center for Global Development.

Retrieved from http://liveweb.archive.org/http://www.cgdev.org/doc/eval%20gap/draft9.15.05web.pdf

Ezemenari, K., Rudqvist, A., & Subbarao, K. (1999). *Impact evaluation: A note on concepts and methods.* Washington, DC: Poverty Reduction and Economic Management Network. World Bank.

Fiszbein, A., & Schady, N. (2009). *Conditional cash transfers. Reducing present and future poverty.* World Bank Policy Research Report, Washington, DC: The World Bank.

Foresti, M. (2007). *A comparative study of evaluation policies and practices in development agencies* (Series Notes methodologiques No. 1). Agence Française de Developpement.

Gamse, B. C., & Jacob, R. T. (2008). *Reading First impact study: Final report.* Cambridge, MA: Abt Associates. Retrieved from http://ies.ed.gov/ncee/pdf/20094038.pdf.

International Food Policy Research Institute (1999). *Progresa. Evaluación de resultados del Programa de Educación, Salud y Alimentación PROGRESA.* Washington, DC: IFPRI. Retrieved from http://evaluacion.oportunidades.gob.mx:8010/es/docs/docs_eval_1999.php.

International Food Policy Research Institute (2000). *Metodología. Evaluación de PROGRESA.* Washington, DC: IFPRI. Retrieved from http://evaluacion.oportunidades.gob.mx:8010/441c7c1a3d30adf64e0e724174a9d527/impacto/2000/ifpri_2000_metodologia.pdf.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., et al. (June 2010). *Single-case design technical documentation. Version 1.0.* What Works Clearinghouse.

Krueger, A. B. (2000). *The effect of attending a small class in the early grades on college-test taking and middle school test results: evidence from project STAR* (Working Paper 7656). Cambridge, MA: National Bureau of Economic Research.

Lachenmann, G. (1988). *Sozio-kulturelle Bedingungen und Wirkungen in der Entwicklungszusammenarbeit.* Berlin, Germany: Deutsches Institut für Entwicklungspolitik.

Leeuw, F. L., & Vaessen, J. (2009). *Impact evaluation and development: NONIE guidance on impact evaluation.*

Levine, R., & Kinder, M. (2004). *Millions saved: Proven successes in global health.* Washington, DC: Center for Global Development.

Levy, S. (1991). *Poverty alleviation in Mexico.* Policy Research and External Affair Working Paper No. 679. Washington, DC: World Bank.

Levy, S., & Rodríguez, E. (2004). *Economic crisis, political transitions, and poverty policy reform: Mexico's Progresa-Oportunidades program.* Policy Dialogue Series. Washington, DC: Inter-American Development Bank.

National Institute of Child Health and Human Development (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of*

the scientific research literature on reading and its implications for reading instruction (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

National Institute of Child Health and Human Development (2000). *Report of the National Reading Panel. Teaching children to read: an evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups* (NIH Publication No. 00-4754). Washington, DC: U.S. Government Printing Office.

Lustig, N. (2011). Scholars who became practitioners. The influence of research on design, evaluation, and political survival of Mexico's antipoverty program Progresa/Oportunidades. *Center for Global Development's Working Papers*, (263).

Maluccio, J. A. (2005). *Coping with the "coffee crisis" in Central America. The role of the Nicaraguan Red de Protection Social.* Washington, DC: International Food Policy Research Institute.

Maluccio, J. A., & Flores, R. (2005). *Impact evaluation of a conditional cash transfer program. The Nicaraguan Red de Proteccion Social.* Washington, DC: International Food Policy Research Institute.

Michigan Department of Education (July 1, 2002). *Making Reading First in Michigan.* Retrieved from http://www.mireadingfirst.org/downloads/mdegrant.pdf.

Miguel, E., & Kremer, M. (2001). *The illusion of sustainability: Comparing free provision of deworming drugs and other "sustainable" approaches in Kenya.* Cambridge, MA: Jameel Poverty Action Lab.

Miguel, E., & Kremer, M. (2003). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Working Paper*.

Mishes, L., & Rothstein, R. (Eds.) (2002). *The class size debate.* Washington, DC: Economic Policy Institute.

National Research Council (1969). *Drug efficacy study. Final report.* Washington, DC: National Academy of Sciences.

Nelson, Richard R. (2001). *On the uneven evolution of human know-how*. ISERP Working Paper 01-05.

OECD Development Assistance Committee (2002). *Glossary of key terms in evaluation and results based management,* from "http://www.oecd.org/glossary/".

OECD Development Assistance Committee (2010). *Quality standards for development evaluation: DAC Guidelines and Reference Series.*

NONIE statement (January 5, 2008). *NONIE statement on impact evaluation (draft).* Washington DC: Network of Networks on Impact Evaluation.

NONIE Subgroup 1 (January 2008). *Experimental and quasi-experimental approaches to impact evaluation.* Washington, DC: Network of Networks on Impact Evaluation.

NONIE Subgroup 2 (January 2008)*: NONIE impact evaluation guidance.* Washington, DC: Network of Networks on Impact Evaluation.

Office of Inspector General (2006). *The Reading First program's grant application process. Final inspection report.* ED-OIG/A03G0006. Washington, DC: U.S. Department of Education.

Office of Inspector General (2007). *The Department's administration of selected aspects of the Reading First program. Final audit report.* ED-OIG/I13-F0017. Washington, DC: U.S. Department of Education.

Reveiz, L., Cardona, A. F., & Ospina, E. G. (2007). *Antibiotics for acute laryngitis in adults. Cochrane Database of Systematic Reviews.*

Paxson, C., & Schady, N. (2007). *Does money matter? The effects of cash transfers on child health and development in rural Ecuador.* Washington, DC: World Bank Policy Research Working Paper 4226.

Pearson, G. (2007). *Battling big pharma: Uncover bias in clinical trials: drug research should be objective, but when it's funded by manufacturers, study results are often predictably rosy. Look for biases that taint clinical trials.* Retrieved from http://www.thefreelibrary.com/.

Rosenberg, L. (July 24, 2007). Project evaluation and the project appraisal reporting system. Washington, DC: U.S. Agency for International Development.

Roodman, D. (2007). *Macro aid effectiveness research. A guide for the perplexed: Working Paper 137.* Washington, DC: Center for Global Development.

Sherman, L. W., Gottfredson, D. C., MacKenzie, D. L., Eck, J., Reuter, P., & Bushway, S. D. (1998). *Preventing crime: What works, what doesn't, what's promising.* Washington, DC: National Institute of Justice.

Snow, C., Burns, S. M., & Griffin, (Eds.) (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academy Press.

Spinks, A., Glasziou, P., & Del Mar, C. B. (2006). *Antibiotics for sore throat. Cochrane Database of Systematic Reviews.*

Stecher, B. M., & Bohrnstedt, G. W. (2002). *Class size reduction in California. Summary of findings from 1999-00 and 2000–01.* CSR Research Consortium.

The National Commission on Excellence in Education (1983). *A nation at risk: The imperative for educational reform. An open letter to the American people: A report to the nation and the Secretary of Education United States Department of Education.*

U.S. Government Accountability Office (2007). *Reading First: States report improvements in reading instruction, but additional procedures would clarify education's role in ensuring proper implementation by states.* Washington, DC: GAO.

U.S. Government Accountability Office (July 2010). *Improved dissemination and timely product release would enhance the usefulness of the What Works Clearinghouse.* Washington, DC: GAO.

U.S. Agency for International Development (January 2011). *USAID evaluation policy. Learning from experience.* Washington, DC: USAID.

U.S. Department of Education (2002). *Guidance for the Reading First program.* Washington, DC: USDOE.

White, H. (2009). *Some reflections on the current debates in impact evaluation.* International Initiative for Impact Evaluation Working Paper 1. New Delhi, India.

White, H., & Barbu, A. (2006). *Impact evaluation: the experience of the independent evaluation group of the World Bank.* Washington, DC: World Bank.

Yatvin, J. (2000). *Minority view. Report of the National Reading Panel: Teaching children to read. Report of the subgroups.* Bethesda, MD: National Institutes of Health, National Institute of Child Health and Human Development.

9. Journal articles

Agerholm, M. (1960). Live polio vaccine. *British Medical Journal, 1*(5177), 966–967.

Allen, L. H. (1995). Malnutrition and human function: A comparison of conclusions from the INCAP and nutrition CRSP studies. *The Journal of Nutrition, 125*(4), 1119S-1126S.

Armitage, P. (1995). Before and after Bradford Hill: Some trends in medical statistics. *Journal of the Royal Statistical Society. Series A (Statistics in Society), 158*(1), 143–153, from http://www.jstor.org/stable/2983408.

Armitage, P. (2003). Fisher, Bradford Hill, and randomization. *International Journal of Epidemiology, 32*(6), 925–928, from 10.1093/ije/dyg286 / http://ije.oxfordjournals.org/content/32/6/925.short.

Ashih, H. W. (2009). Atypicals for non-psychotic disorders. *The Carlat Psychiatry Report, 3*(1-6).

Bamberger, M., & White, H. (2007). Using strong evaluation designs in developing countries: Experience and challenges. *Journal of MultiDisciplinary Evaluation, 4*(8), 58–72.

Behrman, J. R. (2009). Nutritional supplementation in girls influences the growth of their children: prospective study in Guatemala. *American Journal of Clinical Nutrition, 90*, 1372–1379.

Biometrika (1901). Editorial: The scope of Biometrika. *Biometrika, 1*(1), 1–2. Birk, S. M. (2005). Application of network analysis in evaluating knowledge capacity. *New Directions for Evaluation, 2005*(107), 69–79.

Bloom, B. (1984a). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4–16.

Bloom, B. (1984b). The search for methods of group instruction as effective as one-to-one tutoring. *Educational Leadership*, 4–17.

Bluhm, R. (2005). From hierarchy to network. A richer view of evidence for evidence-based medicine. *Perspectives in Biology and Medicine, 48*(4), 535–547.

Bourguignon, F., & White, H. (2007). Aid Effectiveness—Opening the Black Box. *American Economic Review, 97*(2), 316–332.

British Medical Journal (1913). Organization of medical research. *British Medical Journal, 1*(2739), 1381–1382.

British Medical Journal (1955). Poliomyelitis vaccine. American trials in 1954, *1, 4921* (April 30, 1955), 1083–1086.

Brody, H., Miller, F. G., & Bogdan-Lovis, E. (2005). Evidence-based medicine. Watching out for its friends. *Perspectives in Biology and Medicine, 48*(4), 570–584.

Brownlee, K. A. (1955). Statistics of the 1954 Polio vaccine trials. *Journal of the American Statistical Association, 50*(272), 1005–1013.

Carlat Psychiatry Report (2009). *Typical antipsychotics. A brief review.*

Cartwright, N. (2007). Are RCTs the gold standard? *BioSocieties, 2*(01), 11–20.

Cartwright, N. (2009). Evidence-based policy: What's to be done about relevance. *Philosophical Studies, 143*(1), 127–136.

Cartwright, N., & Munro, E. (2010). The limitations of randomized controlled trials in predicting effectiveness. *Journal of Evaluation in Clinical Practice, 16*(2), 260–266.

Caspari, A. (2008). (Rigorous) Impact Evaluations – Eine nicht nur für die Entwicklungszusammenarbeit relevante inter-nationale Diskussion. *Zeitschrift für Evaluation, 7*(1).

Caspari, A. (2009). ‚Rigorose' Wirkungsevaluation – methodische und konzeptionelle Ansätze der Wirkungsmessung in der Entwicklungszusammenarbeit. *Zeitschrift für Evaluation, 8*(2).

Chalmers, I. (2001). Comparing like with like: some historical milestones in the evolution of methods to create unbiased comparison groups in therapeutic experiments. *International Journal of Epidemiology, 30*, 1156–1164.

Chalmers, I. (2003). Fisher and Bradford Hill: theory and pragmatism? *International Journal of Epidemiology, 32*(6), 922–924.

Chambers, R. (1994). The origins and practice of participatory rural appraisal. *World Development, 22*(7), 953–969.

Conway, H., & Clancy, C. (2009). Comparative-Effectiveness Research: Implications of the Federal Coordinating Council's report. *New England Journal of Medicine, 361*(4), 328–330.

Cook, T. D. (2006). What is special about the role of experiments in contemporary educational research: Putting the "gold standard" rhetoric into perspective. *Journal of MultiDisciplinary Evaluation, 6*, 1–7.

Cuban, L. (1988). A fundamental puzzle of school reform. *The Phi Delta Kappan, 69*(5), 340–344.

DeAngelis, K. J. (2003). Class size and student achievement. Lessons from California and Tennessee. *Illinois Education Research Council. Issues in Education*, 1–2.

Denes, C. A. (2003). Bolsa Escola: Redefining poverty and development in Brazil. *International Education Journal, 4*(2), 137–147.

Eisenhart, M., & Towne, L. (2003). Contestation and change in national policy on "scientifically based" education research. *Educational Researcher, 32*(31), 32–37.

Evidence-Based Medicine Work Group (1992). Evidence-based medicine. A new approach to teaching the practice of medicine. *Journal of the American Medical Association, 268*(17), 2420–2425.

Finn, J. D., & Achilles, C. M. (1990). Answers and questions about class size: A statewide experiment. *American Educational Research Journal, 27*, 557–577.

Fisher, R. A. (1918). The causes of human variability. *Eugenics Review, 10*, 213–220.

Fisher, R. A. (1921). Studies in crop variation. An examination of the yield of dressed grain from Broadbalk. *Journal of Agricultural Science, 11*, 107–135.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain, 33*, 503–513.

Fisher, R. A. (1933). The contributions of Rothamsted to the development of the science of statistics. *Annual Report of the Rothamsted Experimental Station*, 43–50.

Freeman, R. (1993). The role of p-values in analysing trial results. *Statistics in Medicine, 12*, 1443–1452.

Friedberg, M., Saffran, B., Stinson, T. J., Nelson, W., & Bennett, C. L. (1999). Evaluation of conflict of interest in economic analyses of new drugs used in oncology. *Journal of the American Medical Association*, (282), 1453–1457.

Galton, F. (1901). Biometry. *Biometrika, 1*(1), 7–10.

Garan, E. M. (2001). What does the report of the National Reading Panel really tell us about teaching phonics? *Language Arts, 79*(1), 61–70.

Glass, G. V., & Smith, M. L. (1979). Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis, 1*(1), 2–16.

Habicht, J.-P., Martorell, R., & Rivera, J. A. (1995). Nutritional impact of supplementation in the INCAP longitudinal study: Analytic strategies and inferences. *The Journal of Nutrition*, 1042S–1050S.

Hanushek, E. A. (1999). Some findings from an independent investigation of the Tennessee STAR experiment and from other investigations of class size effects. *Educational Evaluation and Policy Analysis, 21*(2), 143–163.

Hardy, A., & Magnello, M. E. (2002). Statistical methods in epidemiology: Karl Pearson, Ronald Ross, Major Greenwood and Austin Bradford Hill, 1900–1945. *Zeitschrift für Präventivmedizin*, 80–89.

Hawe, P., Shiell, A., & Riley, T. (2004). Complex interventions: how "out of control" can a randomised controlled trial be? *British Medical Journal, 328*, 1561–1563.

Hilbe, W. (2010). Changing paradigms in clinical trials. *Magazine of European Medical Oncology, 3*(1), 1–2.

Hill, A. B. (1990). Memories of the British streptomycin trial in tuberculosis: The first randomized clinical trial. *Controlled Clinical Trials, 11*(2), 77–79.

House, E. R. (2008). Blowback: consequences of evaluation for evaluation. *American Journal of Evaluation, 29*(4), 416–426.

House, E. R. (2011). Conflict of interest and Campbellian validity. *New Directions for Evaluation, 2011*(130), 69–80.

Hoxby, C. M. (2000). The effect of class size on student achievement: New evidence from population variation. *The Quarterly Journal of Economics*, 1239–1285.

Hubbard, R., & Lindsay, R. M. (2008). Why P values are not a useful measure of evidence in statistical significance testing. *Theory and Psychology, 18*(1), 69–88.

Kaptchuk, T. (1998). Powerful placebo: The dark side of the randomized controlled trial. *Lancet, 351*, 1722–1725.

Kaptchuk, T. J. (2003). Effect of interpretive bias on research evidence. *British Medical Journal, 326*, 1453–1455.

Kaptchuk, T. J., & Kerr, C. E. (2004). Commentary: Unbiased divination, unbiased evidence, and the patulin clinical trial. *International Journal of Epidemiology, 33*(2), 247–251.

Klees, J. E., & Joines, R. (1997). Occupational health issues in the pharmaceutical research and development process. *Occupational Medicine, 12*(1), 5–27.

Kooreman, (2000). The labeling effect of a child benefit system. *American Economic Review*, 571–583.

Koprowski, H. (1960). Historical aspects of the development of live virus vaccine In poliomyelitis. *British Medical Journal, 2*(5192), 85–91.

Lancet (2004). Depressing research, *363*, 1335.

Leeuw, F. L. (2005). Trends and developments in program evaluation in general and criminal justice programs in particular. *European Journal on Criminal Policy and Research, 11*, 233–258.

Luccisano, L. (2004). Mexico's Progresa program (1997-2000). An example of neo-liberal poverty alleviation programs concerned with gender, human capital development, responsibility and choice. *Journal of Poverty: Innovations on Social, Political, and Economic Inequalities, 8*(4), 31–57.

Lexchin, J., Bero, L. A., Djulbegovic, B., & Clark, O. (2003). Pharmaceutical industry sponsorship and research outcome and quality: Systematic review. *British Medical Journal, 326*, 1167–1170.

Martorell, R. (1995b). Results and implications of the INCAP follow-up study. *The Journal of Nutrition, 125*(4), 1127S-1138S.

Martorell, R., Habicht, J.-P., & Rivera, J. A. (1996). History and design of the INCAP longitudinal study (1969–77) and its follow-up (1988–89). *The Journal of Nutrition*, (125), 1027S-1041S.

Mayne, J. (2001). Addressing attribution through contribution analysis. Using performance measures sensibly. *Canadian Journal of Program Evaluation, 16*(1), 1–24.

Mayoux, L., & Chambers, R. (2005). Reversing the paradigm: quantification, participatory methods and pro-poor impact assessment. *Journal of International Development, 17*(2), 271–298.

Medical Research Council (1948). Streptomycin treatment of pulmonary tuberculosis: A medical research council investigation. *British Medical Journal, 2*(4582), 769–782.

Meldrum, M. L. (1998). "A calculated risk": The Salk polio vaccine field trials of 1954. *British Medical Journal, 317*(7167), 1233–1236.

Meldrum, M. L. (2000). A brief history of the randomized controlled trial. From oranges and lemons to the gold standard. *Hematology/Oncology Clinics of North America, 14*(4), 745–760.

Miskel, C. G., & Song, M. (2004). Passing Reading First: Prominence and processes in an elite policy network. *Educational Evaluation and Policy Analysis, 26*(6), 89–109.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children. Critical Issues for Children and Youths, 5*(2), 113–127.

Ritter, G. W., & Boruch, R. F. (1999). The political and institutional origins of a randomized controlled trial on elementary school class size: Tennessee's project STAR. *Educational Evaluation and Policy Analysis, 21*(2), 111–125.

Rogers, P. (2008). Using programme theory to evaluate complicated and complex aspects in interventions. *Evaluation, 14*(1), 29–48.

Rosenberg, W., & Donald, A. (1995). Evidence based medicine: an approach to clinical problem-solving. *BMJ, 310*(6987), 1122–1126.

Rossi, P. H. (1978). Issues in the evaluation of human service delivery. *Evaluation Quarterly, 2*, 573–599.

Straus, S. E., & McAlister, F. A. (2000). Evidence-based medicine: a commentary on common criticisms. *Canadian Medical Association Journal, 163*(7), 837–841.

Schulz, (2004). School subsidies for the poor. Evaluating the Mexican Progresa poverty program. *Journal of Development Economics, 74*(1), 199–250.

Scrimshaw, N. S. (2010). History and early development of INCAP. *The Journal of Nutrition, 140*(2), 394–396.

Scriven, M. (2008). A Summative Evaluation of RCT Methodology: & An Alternative Approach to Causal Research. *Journal of MultiDisciplinary Evaluation, 9*(5), 11–24.

Simon, S. D. (2001). Is the randomized clinical trial the gold standard of research? *Journal of Andrology, 22*, 938–943.

Tanenbaum, S. J. (2009). Comparative effectiveness research: evidence-based medicine meets health care reform in the USA. *Journal of Evaluation in Clinical Practice, 15*(6), 976–984.

Thase, M. E., Macfadden, W., Weisler, R. H., Chang, W., Paulsson, B., Khan, A., et al. (2006). Efficacy of Quetiapine monotherapy in bipolar I and II depression: A double-blind, placebo-controlled study (The BOLDER II Study). *Journal of Clinical Psychopharmacology, 26*(6), 600-609.

Valier, H., & Timmermann, C. (2008). Clinical trials and the reorganization of the medical research in post-Second World War Britain. *Medical History, 52*, 493–510.

Will, C. M. (2007). The Alchemy of Clinical Trials. *BioSocieties, 2*(01), 85–99.

Yates, F., & Mather, K. (1963). Ronald Aylmer Fisher. *Biographical Memoirs of Fellows of the Royal Society of London, 9*, 91–120.

White, H. (2010). A contribution to current debates in impact evaluation. *Evaluation, 16*(2), 153–164.

Whitehurst, (Russ) Grover (2004). Determining 'what works' - An interview with Dr. Grover 'Russ' Whitehurst. *T.H.E. Journal*, January 1, 2004. Retrieved from http://thejournal.com/articles/2004/01/01/determining-what-works--an-interview-with-dr-grover-russ-whitehurst.aspx?sc_lang=en.

Yoshioka, A. (1998). Use of randomisation in the Medical Research Council's clinical trial of streptomycin in pulmonary tuberculosis in the 1940s. *British Medical Journal, 317*, 1220–1223.

Yatvin, J. (2002). Babes in the woods: The wanderings of the National Reading Panel. *Phi Delta Kappan, 83*(5), 364–369.

10. Monographs and book chapters

Banerjee, A. V., & Duflo, E. (2011). *Poor economics: A radical rethinking of the way to fight global poverty*. New York, NY: Public Affairs.

Boruch, R., Moya, D. de, & Brooke, S. (2002). The importance of randomized field trials in education and related areas. In F. Mosteller & R. F. Boruch (Eds.), *Evidence matters. Randomized trials in education research*, 50–79. Washington DC: Brookings Institution Press.

Brookings Institution (1966). *The Brookings Institution: A fifty year Affairs history.* Washington, DC: Brookings.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand McNally.

Campbell, D. T., & Stanley, J. C. (1967). *Experimental and quasi-experimental designs for research* (2. print.). Boston: Houghton Mifflin Company.

Cochrane, A. L. (1971). *Effectiveness and efficiency. Random reflections on health services.* London: Nuffield Provincial Hospitals Trust.

Coles, G. (2003). *Reading the naked truth. Literacy, legislation, and lies.* Portsmouth NH: Heinemann.

Cook, T. D., & DeMets, D. L. (2008). *Introduction to statistical methods for clinical trials. Chapman & Hall/CRC texts in statistical science series.* Boca Raton, FL: Chapman & Hall/CRC.

Dabelstein, N., & Rebien, C. C. (2002). Evaluation of development assistance: Its start, progress, and current challenges. In J.-E. Furubo, R. C. Rist, & R. Sandahl (Eds.), *International Atlas of Evaluation* (pp. 393–405). New Brunswick, N.Y.: Transaction.

Dilthey, W. (1990). Abhandlungen zur Grundlegung der Geisteswissenschaften. In *Gesammelte Schriften* (8th ed.). Göttingen: Teubner; Vandenhoeck & Ruprecht.

Dilthey, W. (2002). Studies toward the foundation of the human sciences. In R. A. Makkreel & J. Scanlon (Eds.), Wilhelm Dilthey. Selected works. Princeton, NJ: Princeton University Press.

Duflo, E., & Kremer, M. (2005). Use of randomization in the evaluation of development effectiveness. In G. Pitman, O. Feinstein, & G. Ingram (Eds.), *Evaluating development effectiveness,* 205–232. New Brunswick, NJ: Transaction Publishers.

Durkheim (1982). *The Rules of Sociological Method and Selected Texts on Sociology and its Methods.* London: Macmillan Press 1901.

Easterly, W. (2006). *The white man's burden: Why the West's efforts to aid the rest have done so much ill and so little good.* New York, N.Y.: Penguin Press.

Emerson, R. M., Retz, R. I., & Shaw, L. L. (1995). *Writing ethnographic fieldnotes.* Chicago: University of Chicago.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh: Oliver & Boyd.

Fisher, R. A. (1932). *The social selection of human fertility: The Herbert Spencer Lecture.* Oxford: Clarendon Press.

Fisher, R. A. (1935). *The design of experiments:* Reissued in *Statistical Methods, Experimental Design and Scientific Inference.* Oxford: Oxford University Press, 1990.

Fleishman, J. L. (2007). *The foundation: A great American secret: How private wealth is changing the world*. New York: Public Affairs.

Garan, E. M. (2002). *Resisting reading mandates. How to triumph with the truth.* Portsmouth NH: Heinemann.

Gertler, J., Martinez, S., Premand, P., Rawlings, L., & Vermeersch, C. M. J. (2011). *Impact evaluation in practice.* Washington, DC: World Bank.

Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Hawthorne, NY: Aldine De Gruyter.

Glennie, J. (2008). *The trouble with aid why less could mean more for Africa.* London: Zed Books.

Guba, E. G. (1978). *Toward a method of naturalistic inquiry in educational evaluation. Monograph Series in Evaluation:* Center for the Study of Evaluation, University of California.

Greenberg, D., & Shroeder, M. (2004). *The digest of social experiments.* Washington DC: Urban Institute.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation.* Newbury Park: Sage Publications.

Gueron, J. M. (2002). The politics of random assignment: Implementing studies and affecting policy. In F. Mosteller & R. F. Boruch (Eds.), *Evidence matters. Randomized trials in education research* (pp. 15–49). Washington, DC: Brookings Institution Press.

Hanushek, E. A. (2002). Evidence, politics, and the class size debate. In L. Mishes & R. Rothstein (Eds.), *The class size debate,* 37–65. Washington, DC: Economic Policy Institute.

Hecht, A., Babcock, H., & Heymann, D. (2009). *Polio*. Deadly diseases and epidemics. New York: Chelsea House.

Hellstern, G.-M., & Wollmann, H. (1984). Evaluierung und Evaluierungsforschung. Ein Entwicklungsbericht. In G.-M. Hellstern & H. Wollmann (Eds.), *Handbuch zur Evaluierungsforschung,* 17–93. Opladen: Westdeutscher Verlag.

Hill, A. B. (1937). *Principles of medical statistics.* London: Lancet.

King Rice, J. (2002). Making the evidence matter: Implications of the class size research debate for policy makers. In L. Mishes & R. Rothstein (Eds.), *The class size debate,* 89–101. Washington, DC: Economic Policy Institute.

Kuhn, T. S. (1962). *The structure of scientific revolutions.* Chicago: University of Chicago Press.

Martorell, R. (1995a). The INCAP Longitudinal Study (1969-1977) and its follow-up (1988-1989): An overview of results. In N. S. Scrimshaw (Ed.), *Community-based longitudinal nutrition and health studies. Classical examples from Guatemala, Haiti and Mexico* (pp. 125–142). Boston Mass.: International Nutrition Foundation for Developing Countries.

Modarresi, S. (1998). *The Profession of Program Evaluation: Theory and Practice.* Dissertation, Albany.

Oakley, A. (2000). *Experiments in knowing: Gender and method in the social sciences.* New York: New Press.

Oshinsky, D. M. (2005). *Polio. An American story : the crusade that mobilized the nation against the 20th century's most feared disease.* Oxford: Oxford University Press.

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage publications.

Patton, M. Q. (2010). *Developmental evaluation: Applying complexity concepts to enhance innovation.* New York: Guilford Press.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation.* London, Thousand Oaks, New Delhi: Sage Publications.

Pitman, G., Feinstein, O., & Ingram, G. (Eds.) (2005). *Evaluating development effectiveness.* New Brunswick, NJ: Transaction Publishers.

Rawlings, L. (2005). Operational reflections on evaluating development programs. In G. Pitman, O. Feinstein, & G. Ingram (Eds.), *Evaluating development effectiveness* (pp. 193–204). New Brunswick, NJ: Transaction Publishers.

Reyna, V. F. (2004). Why scientific reading research? The importance of evidence in changing educational practice. In P. McCardle & V. Chhambra (Eds.), *The voice of evidence in reading research* (pp. 47–58). Baltimore, MD: Brookes Publishing.

Riddell, R. C. (1987). *Foreign aid reconsidered*. London: Overseas Development Institute.

Rogers, E. M. (1962). *Diffusion of Innovations* (3rd). New York, NY: The Free Press.

Rossi, P. H. (1972). Testing for success and failure in social action. In P. H. Rossi, P. H. Rossi, & W. Williams (Eds.), *Quantitative Studies in Social Relations: Vol. 1. Evaluating Social Programs: Theory, Practice and Politics*, 11–49. New York London: Seminar Press.

Rossi, P. H., Freeman, H. E., & Wright, S. R. (1979). *Evaluation: A Systematic Approach* (Vol. 1). Beverly Hills, CA: Sage Publications.

Ryan, G. W., & Bernard, H. R. (2003). Data management and analysis methods. In N. K. Denzin & Y. S. Lincoln (Eds.), *Collecting and Interpreting Qualitative Materials* (pp. 259–309). Thousand Oaks, CA: Sage Publications.

Schriewer, J. (1988). The method of comparison and the need for externalization. In J. Schriewer & B. Holmes (Eds.), *Theories and methods in comparative education* (pp. 25–83). Frankfurt a.M.: Peter Lang.

Shanahan, T. (2004). Critiques of the National Reading Panel report: Their implications for research, policy, and practice. In P. McCardle & V. Chhambra (Eds.), *The voice of evidence in reading research* (pp. 235–265). Baltimore, MD: Brookes Publishing.

Shavelson, R. J., Towne, L., & Committee On Scientific Principles for Education Research (2002). *Scientific research in education* (2. print.). Washington, DC: National Academy Press.

Sherman, L. W. (2002). *Evidence-based crime prevention.* London: Routledge.

Song, M., Coggshall, J. G., & Miskel, C. G. (2004). Where does policy usually come from and why should we care? In P. McCardle & V. Chhambra (Eds.), *The voice of evidence in reading research* (pp. 445–462). Baltimore, MD: Brookes Publishing.

Sweet, R. W. (2004). The big picture: Where we are nationally on the reading front and how we got here. In P. McCardle & V. Chhambra (Eds.), *The voice of evidence in reading research* (pp. 13–44). Baltimore, MD: Brookes Publishing.

Temin, (1980). *Taking your medicine. Drug regulation in the United States.* Cambridge, MA: Harvard University Press.

Vinovskis, M. A. (2002). Missing in practice? Development and evaluation at the U.S. Department of Education. In F. Mosteller & R. F. Boruch (Eds.), *Evidence matters. Randomized trials in education research* (pp. 120–149). Washington, DC: Brookings Institution Press.

Weiss, C. H. (2004). Rooting for evaluation: A Cliff Notes version of my work. In M. C. Alkin (Ed.), *Evaluation roots. Tracing theorists' views and influences* (pp. 153–168). Thousand Oaks, CA: Sage Publications.

Weitzman, E. A. (2003). Software and qualitative research. In N. K. Denzin & Y. S. Lincoln (Eds.), *Collecting and Interpreting Qualitative Materials* (pp. 310–339). Thousand Oaks, CA: Sage Publications.

Wittmann, W. W. (2011). Principles of symmetry in evaluation research with implications for offender treatment. In T. Bliesener, A. Beelmann, & M. Stemmler (Eds.), Antisocial behavior and crime. Contributions of developmental and evaluation research to prevention and intervention. Cambridge MA: Hogrefe Publications.

Wittmann, W. W., & Klumb, L. (2006). How to fool yourself with experiments in testing theories in psychological research. In R. R. Bootzin & P. E. McKnight (Eds.), APA decade of behavior volumes. Strengthening research methodology. Psychological measurement and evaluation. Washington, DC: American Psychological Association.

VITA

Rahel Christine Kahlert graduated from the humanistic branch of the Gymnasium in Tamsweg, Austria, in 1993. She earned a Master of Theology from the University of Vienna, Austria, in 1998 and a Master of Public Affairs from the University of Texas at Austin in 2001. From 2001 to 2006, Ms. Kahlert worked as an evaluation associate at the Charles A. Dana Center at the University of Texas. Ms. Kahlert also worked as evaluator of higher education in various capacities at the University of Texas at Austin.

Permanent email: rahelck@gmail.com

This dissertation was typed by Rahel Christine Kahlert.